DOCUMENT RESUME

ED 084 641 80 EA 005 611

AUTHOR Womer, Frank B.
TITLE Developing a Large Scale Assessment Program.
INSTITUTION Colorado State Dept. of Education, Denver. Cooperative Accountability Project.; Minnesota State Dept. of Education, St. Paul.
SPONS AGENCY Office of Education (DHEW), Washington, D.C.
PUB DATE 73
NOTE 141p.
AVAILABLE FROM Cooperative Accountability Project, Colorado Department of Education, 1362 Lincoln Street, Denver, CO 80203

EDRS PRICE MF-$0.65 HC-$6.58
DESCRIPTORS Developmental Tasks; *Educational Accountability; *Elementary Schools; Evaluation Criteria; *Measurement; *Program Evaluation; *Secondary Schools; State Programs; State Standards; Task Analysis; Testing Programs
IDENTIFIERS *Assessment; Elementary Secondary Education Act Title V; ESEA Title V; Task Development; Task Planning

ABSTRACT

Education has become one of America's biggest industries. As a result, the public wants some evidence concerning the quality of the educational program. This booklet was written to delineate the role of assessment as an important aspect of a comprehensive accountability system for the nation's elementary and secondary schools. It describes what assessment is, provides an overview of assessment, and tells how to plan and develop tasks in an assessment program. Several State assessment programs are alsc described and a checklist for evaluating an assessment program is provided. (JF)

FILMED FROM BEST AVAILABLE COPY

U.S. DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED 084641

DEVELOPING A LARGE SCALE ASSESSMENT PROGRAM

by

Frank B. Womer
Professor of Education
The University of Michigan

Arranged and Edited by

The Project Staff
Arthur R. Olson, Director
Arthur P. Ludka, Assistant Director

Cooperative Accountability Project
Denver, Colorado
1973

EA 005 611

Program Administrators

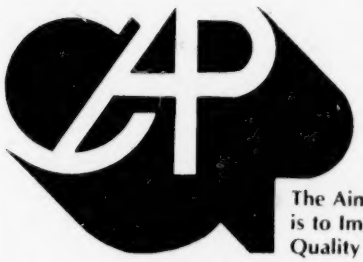
Harry Phillips
Dexter Magers
U.S. Office of Education
Bureau of Elementary and Secondary Education
State Agency Cooperation

Minnesota Coordinator

John W. Adams
Director, State Educational Assessment
Minnesota Department of Education

Copies May Be Obtained From:

COOPERATIVE ACCOUNTABILITY PROJECT
Colorado Department of Education
1362 Lincoln Street
Denver, Colorado 80203



The Aim of Accountability is to Improve the Quality of Education.

the operations board

CAP is guided by an Operations Board consisting of the Chief State School Officers of the cooperating states or their designees.

Colorado Donald D. Woodington, Commissioner, Chairman of the CAP Operations Board; and Stanley A. Leftwich, Assistant Commissioner, Department of Education.

Florida Floyd T. Christian, Commissioner; and James C. Impara, Administrator, Educational Accountability, Department of Education.

Maryland James A. Sensenbaugh, State Superintendent; and Wm. D. Griffith, Research and Pupil Services, Department of Education.

Michigan John W. Porter, Superintendent of Public Instruction; and C. Philip Kearney, Associate Superintendent, Department of Education.

Minnesota Howard B. Casmey, Commissioner; and John W. Adams, Director, State Educational Assessment, Department of Education.

Oregon Dale P. Parnell, Superintendent of Public Instruction; and R. B. Clemmer, Coordinator, Division of Planning, Development and Evaluation, Oregon Board of Education.

Wisconsin Barbara Thompson, State Superintendent; and Archie A. Buchmiller, Assistant State Superintendent, Department of Public Instruction.

U.S. Office of Education Dexter A. Magers, Education Program Specialist and CAP Coordinator.

CAP is a seven-state, thirty-nine month project initiated in April 1972 and financed by funds provided under the Elementary and Secondary Education Act of 1965 (Public Law 89-19, Title V, Section 505, as amended) with Colorado as the Administering State.

FOREWORD

Developing a Large Scale Assessment Program was written to delineate the role of assessment as an important aspect of a comprehensive accountability system for the nation's elementary and secondary schools.

Americans currently are spending well over $100 billion for all forms of education, public and private. This means that education has become one of America's biggest industries. As a result, the public wants some evidence concerning the quality of the educational program.

In a society where education is everybody's business, the adequacy of that system becomes everybody's business. If Americans today were satisfied with their society, if they were united in their perception of social priorities, and if they felt reasonably secure about their power to shape and control the future, there would be less concern today about the concept of accountability. Therefore, it is not surprising that education should come under increasing scrutiny from those on the left, the right, and in the middle, who feel a need to know what the schools are accomplishing, why they are accomplishing or not accomplishing, and how their accomplishments can be improved. Citizens, including professional educators, now see assessment as a means to "get a handl " on what has been happening and on what should happen in terms of student performance.

Just as the world is in a state of increasingly rapid change, so is the school. A generation ago only three groups of people were seriously involved in designing the educational program: the curriculum makers, the school administrators, and certain instructional theorists. What a contrast with the situation now! These people are still involved, but they have been joined by several other important groups: teachers, parents and other concerned citizens, government officials, evaluators, planners, journalists, tax groups, and professional teachers' organizations. *Developing a Large Scale Assessment Program* is addressed to all these groups, as well as to the scholarly community interested in education.

For all these people, this publication offers an assessment plan to think about and set in motion at the state and local school district levels. It suggests specific action steps that should be taken to determine what and how much the students are

learning along with the procedures to be followed to bring about desired improvements.

The Cooperative Accountability Project gratefully acknowledges the leadership and assistance of those members of the Minnesota State Department of Education who worked with Dr. Frank B. Womer in outlining the contents of the publication: Commissioner Howard B. Casmey, Gregory J. Waddick, Assistant Commissioner, John W. Adams, Director, State Educational Assessment and several others who gave their time and advice to make *Developing a Large Scale Assessment Program* an important contribution to the literature on assessment in education.

Arthur R. Olson, Director
Cooperative Accountability Project

PRE-TEST

Answer the following three questions to the best of your ability. This is a test of your knowledge of large scale assessment. The results will not be used in any way to grade or evaluate you. The results will be used to help tell us something of the state of knowledge nationally about assessment purposes and procedures. Please use a separate answer sheet.

(True or False) 1. The long range purpose of every large scale educational assessment project is to effect po.itive changes in education.

(choose one) 2. Every state should develop its own state assessment program because:

a. It is educationally fashionable to do so in the 1970's.
b. The National Assessment project has established a pattern that can be adopted or adapted by any state.
c. State legislators will mandate it anyway.
d. It can provide useful information for improved educational decision making.

(One page or less) 3. Outline, in eight to ten steps, the major types of decisions that n ist be made in developing any larg scale assessment. Explain each one in a sentence or two, *or* give one basic assessment principle for each step.

(Ten minutes later)

STOP. *You must stop now.*

If you would like to know your own score, you should read the following monograph. Put your answer sheet aside now and score it after completing this monograph.

Frank B. Womer

TABLE OF CONTENTS

CHAPTER Page

1 WHAT ASSESSMENT IS 1

What Assessment Isn't 2
Assumptions Underlying Assessment 5
Defining Assessment Specifically 7

2 AN ASSESSMENT OVERVIEW 9

Initiating Assessment 9
Determining General Outcomes 12
A Hypothetical Situation 18
Interpretation of the Situation 19
Considering Side Effects 22

3 PLANNING TASKS IN ASSESSMENT 25

Determining Administrative Structure 25
Determining Types of Assessment Outcomes 32
Determining Types of Test or Survey Instruments . . . 34
Determining Reporting Categories 37
Determining Population 40
Determining Whether Sampling Is To Be Used 41
Determining Data Collection Methodology 43
Determining Data Analysis Strategy 46
Determining General Reporting Policies 49
Determining Costs and Alternatives 50
Determining Time Schedule 53

4 DEVELOPMENT TASKS IN ASSESSMENT 54

Developing Assessment Objectives 54
Using Existing Instruments 57
Developing Instruments 59
Selecting a Sample 66
Collecting Data 71
Processing Results 73
Reporting 77

Chapter Page

5 EVALUATING ASSESSMENT 88

Checkpoints for a Quality Assessment Program 88

6 ASSESSMENT IN OPERATION 91

State Assessment Questionnaire 91
Minnesota Statewide Educational Assessment Program . 94
Texas Education Agency Assessment Program 100
Pennsylvania's Educational Quality Assessment Program 109
Florida State Assessment Program 116
Michigan Educational Assessment Program 126

GENERAL REFERENCES 132

Chapter 1

WHAT ASSESSMENT IS

Educational assessment has been defined, or at least explained, in a great variety of ways. If one searches for some common elements among the definitions and explanations offered, there seem to be at least two that emerge. First, assessment is one process of information gathering. Second, the general purpose of assessment is to use the information that it generates in improved educational decision making. A third element seems to be almost a common ingredient of assessment, and probably should be added to any definition. It is the emphasis on outcomes of education, the knowledge, skills and attitudes that students acquire and develop. One can put this together into a statement that most assessment specialists probably would accept: *educational assessment is a program to improve educational decision making by securing information about the outcomes of education.*

Clearly such a definition is so general that it can encompass a wide variety of assessment plans. This has been true from the very beginnings of the assessment movement in the early 1960's. At that time, the National Assessment of Educational Progress (NAEP) was established to provide, on a sampling basis, information about educational outcomes at the national and regional levels. At that same time, the Pennsylvania Department of Education began the development of a state assessment plan that was geared to providing information about educational outcomes at the local school district level—with instrumentation and norms provided at the state level. These two assessment plans are quite different in their specific goals, yet both are concerned with improving educational decision making based on information about student outcomes.

Some readers may well object to dating the beginnings of educational assessment in the early 1960's. They can, rightfully, point to the many state testing programs and large city testing programs that were started in the 1930's, and even earlier in a few instances. Such programs clearly gathered information about knowledge and skills of boys and girls, in the expectation that teachers would use the results to improve their own classroom practices. The differences between testing and assessment in most instances are matters of details rather than basic intents. Per-

hape the most basic difference is that tests generally are designed to "rank" students on whatever attribute is being measured, whereas assessment instruments generally are designed to "describe" student performance. The term assessment, moreover, rarely was used in connection with the older information gathering programs. It did begin to receive widespread use with the advent of National Assessment (a clearly different program), the requirements of Titles I and III of the Elementary and Secondary Education Act (ESEA), and other similar influences. Currently the terminology of assessment has become so popular in educational circles that what formerly had been called testing is now being referred to as assessment, along with all of the newer or revised practices that are emerging.

In addition to the three essential elements of assessment (decision making based on information about outcomes), there is a fourth element that many, but not all, assessment definitions include—*investigating change in outcomes over time*. National Assessment has made this goal its major purpose. Some state assessments also are stressing the identification of increments or decrements in knowledge and skills and attitudes over time. But not all large scale assessment projects have made change a focal point, so it probably should be included among the great variety of specific goals of assessment rather than in the definition of assessment itself.

What Assessment Isn't

Educational terminology does not carry the agreed-upon precision of the terminology of some other disciplines. Thus, it does not suffice simply to define a concept such as assessment. In order to feel reasonably sure that a reader understands what an author means in almost any area of education, it often is necessary to explain what a definition does not include. This certainly is true when discussing assessment, a relatively new concept that has inherited meaning from older terminology.

When National Assessment first began receiving national publicity it was welcomed by some educators and condemned by others. While some of the reactions were made by persons who took time to understand the real intent of that project, many other reactions, both positive and negative, were based upon misconceptions of National Assessment's purposes. Because assessment was new and unknown, but traditional state testing pro-

grams were well known, some observers assumed that National Assessment planned to develop an every-pupil testing program that would publicize results for individual school districts. Most educators who made this assumption condemned the project; most lay observers of education who made this assumption praised the project. Both attitudes, based upon an assumed, personal definition of assessment rather than upon the stated purposes of National Assessment, were wrong.

Perhaps *the most serious misconception of what assessment means is to equate it to the concept of accountability*. The crucial difference between the two terms is that accountability carries a stronger aspect of value judgment than is present in assessment. Certainly assessment is a process that feeds into decision making, and decision making is based upon value judgments. But accountability, as a concept, focuses first upon goodness or badness, adequacy of performance or lack of it. It is a broader concept than assessment since it encompasses all of the information gathering of assessment plus information about inputs into the educational process, about methodology of instruction, about philosophy of education, and so on. Accountability represents a process that is systematically applied in the education system. Assessment can be a most valuable tool—an indispensable tool—for educational accountability; but it is a part, not the whole. In like fashion, one may think of traditional standardized testing as a potential tool for assessment, but not its sole resource. The natural progression from testing to assessment to accountability has led too many educators to assume that the reverse sequence is equally valid. *It is not correct to assume that accountability is simply assessment and that assessment is simply testing*, so that initiation of a traditional testing program could be assumed to constitute the complete educational accountability process. Unfortunately too many legislators have made this assumption as they drew up assessment legislation. Unfortunately too many educators have encouraged them in that belief.

Another *misconception is the equating of assessment to causal research*. The lack of "control" of relevant factors differentiates assessment from that type of research. Whenever one systematically gathers information about educational outcomes one will find differences in achievement for differing reference

or reporting groups, e.g., big cities vs. rural areas. It is very tempting to assume causal relationships for differences that appear. But results from an assessment project should not be interpreted that way unless all other possible causal factors have been controlled. Certainly those reference groups that are identified as having different assessment results may well bear some responsibility for the results, but the assessment results alone do not tell us that. Anyone who embarks upon large scale assessment must be prepared to face this potential misinterpretation of results. At the same time, one of the great strengths of assessment is that it does "call attention" to other characteristics that are allied to high or low student achievement. One of the great untapped outcomes of a good assessment program is the spin-off of research projects designed to investigate questions of causality that are raised by assessment results.

Another example of what assessment is not is "needs assessment." Certainly no one has a copyright on the word assessment. Those who work on needs assessments know rather precisely what they are doing and are well aware of the fact that "educational outcomes assessment" is a different process with a different purpose. But many educators and undoubtedly most lay persons do not know the difference. *Needs assessment is concerned primarily with determining or establishing what students should be learning. Assessment (educational outcomes assessment) is concerned primarily with determining what students have learned.* The concepts clearly overlap, since knowing what students have learned can contribute to judging what they should be learning, but the concepts are not synonymous. Perhaps one should attempt to coin a new terminology, such as "product assessment" to differentiate what is meant here from needs assessment. But we have a surfeit of educational terminology now.

And finally, *most assessment projects are not concerned to any great extent with measuring inputs into education*, such as financial support, teacher qualifications, curricula, etc. *The main focus of assessment is upon what has been learned*, not the various conditions of learning. Yet what-has-been-learned is not an isolated element of education. Many of those who have promoted assessment for the last ten years have avoided looking at inputs for fear of unwarranted causal implications, and some assessment projects collect no input information. But that attitude is not universal, and some large scale assessments currently

in operation (Michigan, for example) are collecting information about a variety of educational conditions that may well influence educational outcomes in some fashion. Even in these assessment programs, however, the emphasis is upon the outcomes, not the inputs.

Educational assessment is still in its infancy. There are a great variety of information gathering programs that are being called assessment programs. At this stage in the development of assessment, one must define his terms very carefully and talk about specific programs if he wants to communicate. One must refer specifically to National Assessment or to Michigan Assessment or to Florida Assessment or to Pennsylvania Assessment. Each program is quite different in its purposes and procedures. But each one of them is a tool for improved educational decision making.

Assumptions Underlying Assessment

There are at least two important assumptions underlying any large scale assessment program that need to be considered and accepted before launching into assessment. If either one of the assumptions is rejected it would be wasteful to embark upon an assessment program at all.

One of these is that *important information that is not available elsewhere can be obtained from a projected assessment program.* Embarking on a quality, large scale assessment program is a time consuming and expensive project. Moneys for education are too scarce to be allocated to an endeavor that does not give promise of paying off with important, useful information about education that we do not already have. An initial reaction to this assumption may well be, "But of course it is obvious that we can get information about reading and writing and arithmetic because tests already exist to measure these attributes." Certainly it is obvious that methodologies do exist to gather important information about educational outcomes. The more precise question, then, that must be asked in connection with a specific assessment program is: "Does a methodology exist to produce the specific information that this assessment program has been designed to gather?" In most cognitive areas the answer will be "yes"; but in some areas(affective or psychomotor) the answer may not be positive, or if it is positive the associated

price tag may be too high. In order to develop an assessment program to determine how honest students are, it might be necessary to consider some long range observational procedure that followed students into their homes and elsewhere, and which would use up hundreds of thousands of dollars that could better be spent elsewhere.

Perhaps, then, this first assumption should be rephrased in more practical terms, e.g., important information that is not available elsewhere can be obtained for a cost that we can afford.

> **Principle:** *Those who develop large scale assessment projects must not fool themselves into thinking that a cost effective methodology exists for producing any information they may happen to want about educational outcomes.*

A second assumption underlying any assessment program is that *the information gathered from an assessment program can and will be used to improve educational decision making.* There are two aspects to this assumption, the actual potential utility of important information and the potential acceptance of that information by educational decision makers for their own uses. While these are different assumptions in theory, they overlap so much in practice that it seems wise to consider them as two aspects of the same assumption.

No educational assessment program can be justified solely on the basis of gathering important information. The ultimate justification of assessment is that some legislator or some Chief State School Officer (CSSO) or some teacher or some principal or some other person does look at assessment results carefully and make some decision(s) or take some action(s) that is related to insights gained from that information. *Assessment is an educational tool; it must be used to be justified.*

> **Principle:** *The information to be gathered by an assessment program must be information that educational decision makers do need and want and will use if it is provided to them in a form that they can understand.*

Defining Assessment Specifically

There are in existence today a number of state assessment projects as well as National Assessment. How has each of these defined assessment?

National Assessment has defined assessment, for its purposes, as a look at change in educational outcomes over time on a national basis. It has confined its efforts to those knowledges, skills and attitudes that are accepted nationally as important outcomes of education. It has looked upon its role as a census-taker of the educational pulse from time to time in ten different subject areas. All of its reports prior to sometime in late 1973 or early 1974, however, have dealt with status rather than change. Its first "change" report had to wait until one of the subject areas had been assessed twice. Of necessity it took ten years of effort to translate the idea of National Assessment into results that deal directly and specifically with the basic purpose of the project.

Pennsylvania defined assessment, for its purposes, as the establishment of an assessment "service" to school districts in the state. Ten major goals of education were identified statewide. Then tests were selected or developed that were felt to be measures of the ten goals. State norms were established on a sampling basis. Now these tests are offered to individual school districts who want to test their students in order to compare results to state norms. The Department of Education also offers consultant help to schools after assessment results are returned to them. The focus is on school district results and no statewide information is gathered (since the initial test norming).

Michigan's state assessment was mandated by the state legislature and its "real" purpose was rather fuzzy and certainly transitory. All students in two grades were tested in verbal analogies, reading, language, and arithmetic with results reported back to schools, and comparative school district rankings were eventually published statewide. After the initial assessment was complete, it was decided to tie the allocation of some compensatory funds to the school district results. A change in purpose is now leading Michigan to development of a plan for assessing minimal every-pupil competencies in a variety of subject areas at three or four or more different grade levels. Michigan has not yet settled on a permanent definition of assessment.

Minnesota conducted a pilot assessment program in order to help set the direction for longer range planning. As a result, the following objectives were developed for statewide assessment:

- To determine the level of performance of students in the cognitive, affective, and psychomotor domains.
- To identify variables which are related to student performance.
- To use the results as a guide for the allocation of school resources by educational decision makers.
- To determine the extent to which progress in student performance is being made in Minnesota.

From the above examples, it is apparent that each assessment program takes on its own characteristics in accord with its purposes. As indicated previously, a person must be definitive and talk about specific programs when referring to assessment.

Chapter 2

AN ASSESSMENT OVERVIEW

Assessment is in fashion. Most states either have embarked on statewide assessment or are seriously considering it. How did we get to this point? Who has decided that we must assess? What is the best impetus for initiating an assessment?

Initiating Assessment

Many state departments of education are engaged in assessment because they have been forced to, sometimes willingly, sometimes not so willingly. Federal regulations and directives now require that many funded projects include provisions for regular evaluative assessment, sometimes specifying testing both before and after instruction or other program components.

Michigan has been cited as an example of a legislatively mandated state assessment. The State Department had asked the legislature for planning moneys for state assessment, but instead was required to implement a statewide assessment within months of passage of very specific assessment legislation. Initial legislation in Florida called for the development of assessment procedures which would provide each school district with relevant comparative data and, to the extent possible, be compatible with national procedures for the assessment of progress in education.

Mandated assessments too often suffer from unrealistic expectations on the part of those who initiate them. Some educators and some legislators clearly have become enamored of assessment (testing) as a tool that can (will) provide specific answers to almost any question they may choose to ask about educational outcomes of school districts and school buildings and classes and individual students. In fact, methodologies do exist for providing a great deal of this information in the cognitive domain at least, but the price tag for really fulfilling all of these expectations is far greater than anything yet projected for any assessment project. The non-technician in educational measurement—either inside or outside the educational profession—often expects test results to provide much more than they are capable of providing. National Assessment has spent many millions of dollars in the process of providing good, accurate

output information on the national level only, using sampling approaches. *Good information is costly; poor information isn't worth a nickel.*

In Pennsylvania the impetus for assessment came from the Department of Education. This was also true in Maine, North Carolina, Colorado and Minnesota. It is not at all unusual for a CSSO to see assessment as a tool that will provide him with information useful in planning departmental priorities and activities, or to see assessment as a state service to Local Education Agencies (LEAs), or perhaps, defensively, to see assessment as a means for publicizing educational outcomes within the state. For a great variety of reasons, different state departments of education have initiated state assessment programs.

It may be assumed by some educators that an educational program initiated within the educational establishment will have a much better chance of producing effective results than an externally imposed program. While it probably is true that most educators feel better about projects that are internally generated, it is not obvious that the end products are necessarily superior.

Assessment at the LEA level is still in its infancy, unless one defines traditional standardized testing as one aspect of assessment. Some activities that formerly were thought of as testing now are thought of as assessment, and some test publishers are using assessment terminology to market "tests." *It is much easier in education to change labels than to change products or procedures.* While enologists may be concerned with new wine in old bottles, educators should be concerned about old wine in new bottles. However, the current thrust toward assessment—toward looking more closely at educational outcomes, at specific knowledge and skills—is bound to have an effect on testing at the local school district level. Perhaps it will result more in revised, and renamed, testing programs than in "new" assessment programs.

The initiation of National Assessment was in a slightly different manner. It began as an idea that there was a need to look broadly at educational outcomes across the nation. Sputnik had triggered many changes in American education, one of which was concern about lack of information as to whether schools were or were not producing enough scientists and mathematicians. Frances Keppel, then the United States Com-

missioner of Education, called together an ad-hoc group of educational leaders to discuss the idea of assessment at the national level. John Gardner, Ralph Tyler and James Allen were in that group. The result was a foundation-sponsored independent organization designed to explore in depth the feasibility of a national assessment, and to develop implementation if it were deemed feasible. The end product, today, is an ongoing census of educational outcomes, called National Assessment, supported by the U.S. Office of Education, but operated by the Education Commission of the States. *National Assessment* has been successful. It is *an example of a program initiated with adequate support, with access to some of the best educational thinking in the nation, and with sufficient time to develop a new program.* That formula is hard to beat.

How should large scale assessment be initiated? It should be initiated by men and women who have questions about educational outcomes for which they need answers in order to help themselves and others make better educational decisions. Large scale assessment should be initiated by men and women who are willing to support quality assessment procedures and willing to wait for the development of a quality program. It doesn't really matter whether it is a state legislator, or a CSSO, or a director of research, or a school superintendent, or someone else in a position to influence activities of a State Education Agency (SEA). It is important that the person promote it for what it can do. Of course, it would be unrealistic to expect a state assessment program to be successful without the commitment and active support of the CSSO.

Principle: *It is possible to mandate a state assessment program; it is not possible to mandate a quality state assessment program.*

Principle: *A mini-assessment or pilot assessment, well supported and well executed, probably is the best possible way for a state or other large educational unit to begin assessment. Whether it has been mandated or not is less important to its ultimate utility than is its careful development.*

> **Principle:** *The measure of the success of any large scale assessment will be directly related to the quality of staff that is assigned to the program and the amount of support devoted to it. The locus of initiation of that program will be only peripherally related to its success, if at all.*

Determining General Outcomes

Undoubtedly the first real assessment decision that must be made by any director of a large scale assessment program is simply, "What shall we assess?" It is easy to react quickly to this question and to generate a superficial response that makes sense, but does not form a sound basis for assessment.

A typical reply might be, "We ought to assess reading," or "The basic skills," or, by some people, "We should find out if we are producing good citizens," and so on. The problem with all of these answers it that they make too great a leap and thus bypass a series of very specific, but essential decisions that must be made prior to any effective choice of subject area or areas.

Before even thinking about reading, or mathematics, or citizenship, or science, a director of a new assessment program should think about who will use the assessment results, what kinds of educational decisions they will make, what types of information they will need, how they will (can) interpret the information provided, and what type of applications really are envisaged for the results. These concerns will be among the most difficult ones he will face. If he concerns himself only with implementation of "some" assessment program, he will eventually face that most difficult assessment question of all, "What has happened as a result of your assessment that has contributed to education?"

First of all it is essential to identify specifically the major audience or audiences for which the assessment is designed. Some of the following questions must be considered:

- Is a state assessment being designed primarily to provide state legislators with information that they need to do a better job of voting on educational bills—perhaps on a bill that would appropriate or reallocate more money for reading improvement?

- Is a state assessment being designed primarily to provide the CSSO and his staff with information that they need to do a better job of providing consultant services to schools—perhaps consultant services in vocational education programs?

- Is a state assessment being designed primarily to provide superintendents of LEAs with information that they need to do a better job of evaluating whether most of their students are mastering basic skills—perhaps those communication skills that are deemed essential for most students?

- Is a state assessment being designed primarily to provide individual teachers with knowledge of whether their students are growing into good citizens, perhaps in the area of tolerance for a diversity of cultural and ethnic backgrounds?

It is not sufficient simply to decide to do an assessment in reading, or arithmetic, or citizenship. It is more important to gather information that will be maximally useful to someone, the legislator, the CSSO, superintendent, the teachers, the pupils, the parents. It is not possible to provide maximally useful information to each of these audiences—unless one is prepared to institute a uniform every-pupil state program of a detailed diagnostic nature at every grade level.

One point of view claims that if information is gathered that is maximally useful to teachers it can be condensed and summarized for successively larger educational units and, thus, will be useful to everyone else also. But the state legislator and CSSO need information that is uniform across the state and must be obtained either from all students or from a random sample of students. Whereas, if each teacher is to maximize his understanding of specific reading and arithmetic problems of students, he must regularly and routinely have evidence from a variety of diagnostic type tests. *An every-pupil diagnostic testing program at a state level would be a multi-million dollar project.*

Local school districts have rarely, if ever, attempted large scale diagnostic testing. Survey testing has been done and can be

done on a widespread basis for a nominal cost. But survey testing provides only generalized information for teacher use. Survey testing can and often does provide adequate information for those educators making decisions at the building, or district, or state level. These are the levels at which most assessment programs are functioning and will be able to function in the foreseeable future.

> **Principle:** *A multi-purpose assessment program that is designed to meet the needs for educational information of persons at all levels of educational decision making runs the risk of not really providing the information that is maximally useful for any one of them.*

Educational decision makers such as state board members, or legislators, or the governor are apt to ask such questions as the following:

- Are we getting our moneys' worth for our educational dollar?
- Are our students learning as much as the students in neighboring states?
- Why can't our students read?

These three questions could lead to quite different information-gathering approaches. To determine the relative effectiveness of the educational dollar, as suggested by the first question, it would be necessary to agree first upon some specific outcomes for which assessment methodology exists, and then to vary educational inputs (including moneys) to see if relationships exist. Such an approach is more of a causal educational research approach than an assessment approach. One might do this in selected school districts, but not statewide at all.

The second question is one to which statewide assessment lends itself very readily. Such a question could be answered rather simply with a small sampling of schools around the state. Comparisons between states or regions are quite easy. Maine has done it very simply by using released National Assessment exercises and comparing its results to those of the nation and the Northeastern states.

The third question might be approached in two steps. First, can our students read or can't they? This requires establishment

of agreed-upon standards of reading, with associated test materials, and then a statewide sample of students to see whether there is a reading problem or not. Following the gathering of this assessment information, there should be a plan of action that would proceed in one direction or another depending upon the results. This would be assessment plus a plan for change.

These three examples suggest gathering information or doing research on a relatively small subset of students. Probably most educational information that will be maximally useful to lay persons concerned with the governance of education at the state level will be of this type, i.e., information that can be gathered relatively easily and relatively inexpensively on a sampling basis (thousands or hundreds of thousands of dollars rather than millions of dollars).

What type of educational questions are asked by the CSSO or members of his staff?

- How might we best award discretionary moneys?
- In which schools can the SEA have the greatest impact for good?
- Where are consultant services most needed?

These three questions suggest that an information gathering assessment might be concerned with differences in attainment between districts. An assessment that wants to gather information that will rank order all districts in some fashion must be designed to gather some information in all districts.

Another set of questions might relate to differences in curricula that are relatively independent of school districts. Such an assessment can be designed to sample from districts, or even school buildings that have different curricular patterns, but need not call for assessment in every district.

What types of educational questions do school superintendents, or principals, or teachers have? Some of the questions might be as follows:

- Are our students developing arithmetic skills as well as they should?

- Will team teaching produce any better results in science?

- Can Karen, or Roger, or Sally, or George read satisfactorily?

These questions focus clearly on student outcomes within an LEA. Before deciding how a state assessment should be designed to provide useful information for LEAs, it should be decided whether the SEA, in fact, wants to take on a task of information gathering for LEA needs. Some states have said yes to this, some have said no. For those SEAs that have designed or planned assessments for LEA needs, there seems to be a tendency to develop a "service" assessment program one that LEAs may elect to participate in or may elect to ignore. This seems very logical if the focus truly is on LEA needs. In that situation, there is no need to gather information that applies to an entire state nor to make district comparisons. But if one decides that there is a need for information on a statewide basis by SEA personnel or legislators, or the state board then an assessment program must either be compulsory for all schools or must use a probability sampling approach.

Again we return to the idea that every large scale assessment should be focused on some relatively specific audience(s) and some relatively specific questions for which information about educational outcomes will be useful. Focusing an assessment program on lay persons' questions does not mean that educators will find the resulting information useless.

> **Principle:** *The most logical starting point for any large scale assessment is to determine its prime audiences and their most crucial needs for student outcomes information. Such a focused program probably will be useful to others outside the prime audience, whereas an unfocused program may not be very useful to anyone.*

After determining the basic ground rules for an assessment, there still are a few more questions and concerns that should be

addressed prior to plunging into the actual development of an assessment plan. The needs of the selected audiences and their basic questions must be translated into general areas of information gathering in preparation for the more specific planning that will come later. As examples, the following can be considered:

- In general, do the questions that have been asked call for information about cognitive outcomes, or affective outcomes, or some combination?
- If cognitive, is there greater concern with skill development, or acquired knowledge, or some combination?
- Should assessment look solely at outcomes or should it also gather some information about inputs into the educational process?
- Will the survey approach to assessment be adequate or will it be necessary to build in some control of selected variables?

Answers to these questions usually can be formulated from the educational questions posed by the major audience for an assessment program. Once these have been established *they form reference points against which later specific program formulation may be checked.* Without answers to such questions, a director of assessment is forced to speculate about the "real" concerns and desires of those who have instigated the assessment program.

It would be overly optimistic to assume that all legislators, or all CSSOs, or all school superintendents, or all teachers could or would formulate their own desires for information about educational outcomes in precise, easily understood questions or statements that could be translated directly into an assessment program—even though that is exactly what is needed. The best that can be hoped for, initially, is a general direction for assessment and a mandate to do it. *What can be done, and should be done, is to formulate,* as precisely as possible, *statements of a few major objectives for assessment* as the director and his staff translate that "general direction" into objectives. These objectives should then be taken back to persons with the legal responsibility for assessment and/or to the major audience(s) for the results to

determine whether their general directions or purposes would be served by these specific assessment objectives.

A Hypothetical Situation

A useful technique in this process of trying to establish concrete purposes for an assessment is to illustrate each objective with an example of what a final report might look like if that objective is used, as a guideline for assessment. Think of this example:

Purpose of Assessment in the State of Limbo:

To determine whether basic skills have been developed by high school graduates.

Goal 1. (from a series of goals)
Seniors in Limbo should be able to compute well enough to handle every-day mathematical problems.

Objective 1. (under Goal 1)
Seniors in Limbo should be able to make change accurately.

Subobjective 1. (under Objective 1, of Goal 1)
When given eight different problems of making change in every day situations (groceries, hardware, clothes, records), ninety-five percent of Limbo's seniors will correctly answer six of the eight questions.

* * * * *

Subobjective 1. (under Objective 1, of Goal 2)
When given six different reading selections from "feature" stories of a newspaper, ninety-five percent of Limbo's seniors will correctly answer four of the six questions.

* * * * *

Sample Reports:

1. Statewide, 97 percent of all Limbo's seniors correctly answered six of eight questions relating to making change in common situations.

2. Statewide, 92 percent of all Limbo's seniors correctly answered four of six questions relating to selecting the best title for short reading selections from newspaper feature articles.

3. In 393 of the 421 school districts in Limbo, ninety-five percent of the seniors did correctly answer six out of eight questions relating to making change in common situations.

4. In 376 of the 421 school districts in Limbo, ninety-five percent of the seniors did correctly answer four of six questions relating to selecting the best title for short reading selections from newspaper feature articles.

5. The Pleasantville school district ranked at the 32nd percentile among all school districts in Limbo on eight questions relating to making change in common situations.

6. The Pleasantville school district ranked at the 14th percentile among all school districts in Limbo on six questions relating to selecting the best title for short reading selections from newspaper feature articles.

7. The Pleasantville school district ranked at the 37th percentile statewide on survey arithmetic questions given to seniors.

8. The Pleasantville school district ranked at the 11th percentile statewide on survey reading questions given to seniors.

9. Sally Wilson got all eight questions correct relating to making change in common situations.

10. George Searls got three of the six questions correct relating to selecting the best title for short reading selections from newspaper feature articles.

11. To– 10,000. (Innumerable hundreds and thousands of examples could be prepared, such as the relation of the results in Limbo to results in other states or to the nation as a whole.)

The reader may be puzzled at the sudden inclusion of sample, hypothetical assessment reports in the middle of a discussion of how to determine the purpose of a large scale assessment. But perhaps the very best way to "force" those *persons who establish policy and purpose* is to give them a series of possible types of reports and then *have them decide which ones provide the type of information that they really want*, and the type of information that has the potential for important decision making.

Interpretation of the Situation

If *Sample Report #1* (above) is the type of information that is desired, it implies at least two things–very specific objectives and questions must be written, and only a relatively small statewide sample of seniors need be tested.

If *Sample Report #8* is the type of information that is desired, it implies that a variety of general reading items may be

purchased or written and that the items should be, on the average, of middle difficulty in order to spread district scores as far apart as possible. It also implies that only in large school districts will sampling of seniors be appropriate.

Sample Report #10 implies many very specific objectives and questions and testing all seniors in Limbo. This would be a very expensive assessment if all important skill objectives were assessed.

If a policy maker wanted the type of information in both *Sample Reports #3 and #7* above he could not have it from one assessment and still maintain top quality information for both types of reports. *Sample Report #3* would be a situation in which the average results from 393 school districts would be condensed within 5 percentage points, from 95 to 100. In order to separate the 393 school districts at all (to rank them), one would have to do the ranking on the basis of differences between numbers like 96.27 and 96.34, for example. Classifying one district as superior to another on such information would be educationally ridiculous even though it might be statistically satisfactory. If one's prime purpose is to rank order school districts (a really questionable purpose in the writer's opinion), then any good standardized achievement test will do quite well.

> **Principle:** *In developing the real goals and objectives of a large scale assessment one needs to know, very specifically, the type of information that the persons in charge of assessment policy really want.*

In addition to focusing attention on very specific potential types of information that can be obtained from large scale assessment, there is another most important question to consider before any decision is made as to whether to proceed into an assessment program itself. What will happen after we have information from an assessment program? For example: If we should find that only 92 percent instead of 95 percent of Limbo's seniors answered some reading questions correctly, so what?

- Does this mean that the legislature will appropriate more money for reading improvement of seniors, or should they?

- Does this mean that the CSSO will hire three more reading consultants, assigned to work on reading materials for seniors, or should he?

- Does this mean that all school superintendents in Limbo should institute a unit on reading improvement at grade 12, or should they?

- Does this mean that 12th grade English teachers will or should spend more time on reading skills than on the modern novel?

Clearly there are no answers to these questions from the assessment results. *Information is important in helping people arrive at answers; information is not the answer, nor is it the action.*

The easiest stance for a director of assessment to take is that other people should use this valuable information that has been supplied by the assessment program. A more constructive approach for an assessment director to take is to point out clearly and loudly that the first educational implications from any assessment results undoubtedly will call for more or a reallocation of money, and people, and greater dissemination of materials and new programs, etc. If assessment results fall short of expectations at any point, improvement can take place only if resources are brought to bear on the deficiency. *Educational change does not happen simply because there seems to be a need for change.*

Commitment to state assessment should include commitment to a program or programs that will be designed to effect positive change. If this commitment is not present, one should seriously question the advisability of putting money and human resources into information gathering—which is what assessment is.

If the Pleasantville school district ranks at the 32nd percentile (Sample 5) on some questions about "making change," so what? The 32nd percentile is in the average band, neither high nor low. Perhaps such a score would be useful for a decision as to whether Pleasantville should or should not receive a visit from a math consultant in the state department, or whether Pleasantville should, or should not receive additional moneys, or extra curriculum materials for math education. But *the percentile ranking alone is not sufficient information even to make such relatively*

simple decisions. One may need to know more about socio-economic conditions in Pleasantville, about average scholastic aptitude, about scores in other math areas (higher or lower), and scores in other subject areas (reading, science, social studies), for example.

It is easy to decide, yes, we need an assessment because it will provide information that will help us make better educational decisions. *It is not easy to design one single assessment program that will give valid, reliable answers to the multitudes of questions that are asked about educational outcomes.* One must limit any self-respecting assessment to a few well-chosen, vital educational questions for which we most need information now. One of the best ways of identifying the most vital concerns and establishing priorities is to furnish people with alternative reports that illustrate what they could get from assessments designed in different ways. It then becomes necessary to secure decisions as to whether a sample report does provide the desired information and to tell policy makers what type of program needs to be developed to provide that information, and how much it will cost. Spending $100,000 to secure information that no one will, in reality, use is a waste of money and time. Spending $500,000 to secure information that plays a vital role in allocating millions of dollars or in upgrading curricular practices may be a real economy.

Considering Side Effects

Before a director of a large scale assessment program begins program planning, he should consider one other general point—what side effects an assessment program may generate. The discussion up to this point has been focused on potential utility of assessment results, but what of potential problems? Such potential problems should not deter the development of an assessment program, but they may very well influence the direction that assessment takes at a given point.

Some of the greatest headaches of an assessment director can come from *adverse publicity.* Certainly any public program will be, and must be, subject to criticism. The criticism that hurts most is that which is based on misunderstanding—either of the role of assessment or of specific results that are produced. On the other hand, *critics* of assessment, who criticize because of beliefs that differ with an assessment program which they do understand,

should be listened to for the potential improvements that they may suggest.

One can be very certain that when assessment results are made public, either as tables of numbers or as written discourses or as verbal remarks, some people will latch onto them and use them as justification of already developed ideas and prejudices. Many others, of course, will read (or listen to) results and will interpret them more wisely. Unfortunately, the majority will tend to ignore them. *The mere use of numbers to report assessment results has problems associated with it.* To most lay persons and, unfortunately, to many educators, numbers are magic symbols. To many they represent, by their very existence, a higher level of accuracy than they may attain. To many, they suggest goodness and badness, whereas they really represent only quantity of something. Judgments of goodness or badness are legitimate when comparing numbers to some standard or standards, but the standards are not always well understood. In one sense a director of an assessment project is faced with publishing results in a language that is clearly a "second" language to many of those who read the results, the language of numbers. *Some criticisms of assessment will stem from sheer misunderstanding of numerical reports.* Anticipating this problem may result in better, more lucid assessment reporting techniques. Potential criticisms should not just be rebutted; they should be examined in advance for good ideas that may result from them.

An attempt to anticipate potential criticism can be helpful in total program development. If it is decided that direct school district comparisons are going to be made, one can anticipate that there will be cries of anguish from representatives of at least some of the districts that rank low. It is inevitable that people who are made to look bad will be unhappy. To the extent that a program can be designed and reporting techniques developed that do not "point the finger" at districts, or that minimize the "pointing," the criticism can be reduced. It should be very apparent that one or more of the objectives of assessment will in fact be implemented by providing rank-order comparative school district results before finalizing a decision to do so.

There will be potential criticisms of assessment methodologies, e.g., every-pupil assessment or matrix sampling. There will be potential criticisms of assessment instruments, e.g., norm-referenced or criterion-referenced. Anticipating as many of these

problems as possible may very well produce more well-thought-out rationales for the many assessment program decisions that must be made.

Chapter 3

PLANNING TASKS IN ASSESSMENT

One of the first tasks that a director of assessment should address himself to is the clear establishment and delineation of lines of authority for decision making. For a state assessment program, the State Board of Education (SBE) and the Chief State School Officer (CSSO) will hold responsibility for all major policy decisions about state assessment. They should be asked to decide the purpose and major goals of assessment and even to approve specific objectives of assessment. The CSSO and his staff probably will not themselves prepare specific statements of purpose and goals for assessment, but they must be the final arbiters of the general direction of assessment within the limits of assessment legislation that may be in force.

In some states, the legislature has produced a law that itself is quite directive and which limits the policy options of the SBE. In other situations, the legislation has been more general and the state department has considerable leeway in determining policy.

Determining Administrative Structure

Because SBE members are rarely experts in educational measurement and only occasionally are professional educators, they must rely on others for professional advice on assessment. The CSSO and his staff are their first source of advice. They may wish to augment that advice by establishing a Policy Committee or an Advisory Committee on State Assessment. The major difference between the two types of committees is the degree to which such a committee is given "authority" to establish policy within general broadly-developed guidelines versus being asked for their "advice." And this potential difference may appear to be greater than it actually is. Most SBEs and CSSOs must retain ultimate veto authority on all assessment decisions, but, in practice, will accept the advice of such a committee most of the time—at least if the committee is both representative and capable.

For practical operational reasons it will be essential to decide early whether an advisory committee is to be advisory to the SBE, to the CSSO, or to the assessment staff. In practice it becomes advisory to whichever persons it meets with (generally the assess-

ment staff), regardless of its official appointment or designation. An official advisory committee probably should be appointed by the Board or the CSSO, while other, ad-hoc advisory groups probably should be constituted by the assessment staff to offer advice on specific areas of assessment.

A general, widely-representative *advisory committee should be informed rather specifically what its role or roles will be.* Such a committee, typically, should be asked to deal with fairly broad assessment questions rather than questions of detail. Such a committee can well address itself to questions of what to assess, how to report it, and whether to sample or do an every-pupil assessment. In contrast, a small ad-hoc advisory group of specialists might deal with specific problems of test development, how to collect or construct questions, how to relate them to objectives, and so on.

Sometimes it appears that a general advisory committee has been created to serve as "window dressing" for assessment, or to gain support for assessment from a wide variety of sources. *Some people resent being asked to serve on a committee when the motivation appears to be primarily public relations.* If the sole purpose of such a committee is to feed information to it about assessment, it would be far better to label it as an ongoing dissemination committee and not pretend to seek its advice. However, it does seem quite legitimate to create an advisory committee, with its own self-selected chairman, and actively seek and welcome its advice while still using the members as a means of improving dissemination about assessment. Such a committee is a natural two-way street for carrying questions, concerns, and recommendations about assessment to the central staff, while also carrying accurate information about assessment back to the constituent groups that may be represented on the committee.

Most assessment programs will benefit from having one large, broadly-representative advisory committee that meets periodically to hear progress reports and to help set general direction of the program. *A committee that does nothing but raise questions and issues and does not really offer advice certainly is less than maximally useful.* Membership on such a committee should include persons from all major constituent groups—lay persons, legislators, school administrators, teachers, educational researchers, curriculum specialists, nonpublic educators, etc. One simple way to constitute such a committee is to ask various existing organizations to

nominate persons for membership, e.g., parent-teacher organization, superintendents' organization, teachers' organizations, chamber of commerce, state legislature, labor organization and so on. In some instances it may be appropriate to ask for the official naming of a person as a representative of a specific organization. In other instances it may seem better to ask for nominations only and leave the final selection to the assessment staff and/or the CSSO. The former method may have a slight advantage from a public image point of view; the latter allows for an attempt to balance the committee with persons with specific points of view that do differ. The latter technique may subject one to criticism of stacking a committee, but not if selections really are made to create diversity. The latter also has the advantage of keeping off the committee someone who has a specific axe to grind and will not be satisfied when his point of view is not accepted. A decision as to how to constitute such a broadly representative committee will of necessity reflect local considerations.

Membership on a smaller working committee (perhaps a Technical Committee) that meets fairly often to discuss specific matters of program development probably is best determined by personal knowledge of the abilities, interests, and work habits of the individual members. On such a committee one wants persons who definitely support the idea of assessment, even though they should represent a diversity of opinions about developing an assessment program. (On a larger advisory committee one may want one or more persons who really are opposed to the idea of assessment and want to minimize its effects.) *Members of a working committee must be willing to devote time regularly to assessment concerns.* They should be persons who will present their opinions forcefully, but also are willing to accept compromise, whether or not they agree with it.

The over-riding *qualities* that should be sought in members of such committees are *intellectual insight*, *common sense*, and *breadth of experience*. Some, but not all members should have had experience with assessment, and/or evaluation, and/or testing. In order to secure the best possible advice from an advisory committee, one must secure the best possible members. Most state departments of education will have many competent potential advisers within reasonable driving distance—only a few may suffer from physical isolation.

Ad-hoc advisory committees should be constituted when-

ever specific expertise is needed that is not present on an ongoing committee, or when particular problems call for a special group of advisers. It may be that the development of a sampling plan for an assessment program will benefit by having it reviewed by a small committee of sampling statisticians combined with persons who have broad assessment knowledge. It may be that an assessment program staff will want the advice of three or four public relations and communications specialists when developing their dissemination plan. One such specialist may or may not be a member of the ongoing advisory committees, but rarely would three or four of them be pulled together except on an ad-hoc basis.

A most important concern that relates to the functioning of advisory committees is the aspect of *what questions to take to a committee for advice*. For the most part the assessment staff should decide where they want and need advice, and that should form the basis for agenda items. Even for a general advisory committee it is unrealistic to expect the committee to generate its own agenda. Yet any general advisory committee should have the freedom of adding items to an agenda if it feels that certain points warrant discussion. Committee members may see potential problems and concerns that staff has not anticipated and should be free to bring them up for discussion.

> **Principle:** *Well chosen advisory groups are an invaluable tool of assessment. They provide the opportunity of focusing a broader spectrum of knowledges and skills on assessment problems than could ever be encompassed within an assessment staff.*

Staff. It certainly is a truism to say that *assessment, like any other program, will rise or fall in direct relation to the quality of the staff that is assigned to or hired for assessment.* The assignment of a person on the SEA staff who is currently under-employed to be director of state assessment is not the way to ensure a quality assessment—propinquity and quality are not synonyms. Certainly a state department may have a staff member who is eminently qualified for the position. If that is so, assessment can get underway more quickly than if an outside person is hired. A current staff member will know the office politics and may be able to get decisions and implement them more quickly than a new staff

member.

Typically, a director of assessment should have training and experience in educational measurement or some other aspect of educational research. A director will be called upon to make some decisions that involve technical expertise. And while it is possible to secure technical expertise in educational measurement from other staff members and from consultants, the person who must make most of the day-to-day assessment decisions needs enough knowledge about assessment to be able to understand alternate courses of action well enough that he can make informed decisions. Ideally, a director of assessment also should have some breadth of experience across several educational areas. Teaching and administrative experience in education will be most helpful. Being able to work well with others is essential, since the assessment process brings the director into working relationships with every area of education as well as with interested lay persons. Of necessity, information gathering about educational outcomes is of vital concern to parents, to legislators, to teachers, to curriculum and subject matter specialists, to school administrators, to boards of education, and to the press. A director of assessment will be working with people from all these backgrounds. He must be able to understand their educational concerns. Invariably a director of assessment will be faced with dissemination problems and associated public relations problems. He must be able to face critics who may be hostile.

It is essential that a director of assessment should be flexible in his thinking and should not be tied to any particular scheme of data gathering. This flexibility is essential if an assessment program is to be fitted into the needs for educational information that are of prime concern within the state. *There is no one best assessment program; there is no specific model that should be adopted intact for a given state.* Principles of good assessment development have emerged, but the application of good principles to the development of a specific state assessment calls for a willingness to weigh all existing options and to seek new options. *Flexibility is a must.*

A director of assessment needs to be realistic about what he and the staff can do in assessment development within given time limits and given budget. *A potential director who asks for a year to do assessment planning is much wiser than one who promises results in May if hired in September.* The latter candidate would be able to deliver some results, of course. But whether they will

really be useful and will be geared to the most important needs for information in that state is highly questionable.

One of the *best attributes* that one can seek in a director of assessment *is the ability to find and secure a capable assessment staff.* This quality is valuable to any administrator, of course, not just to the director of an assessment program. However, because of the short history of large scale assessment, there is no large body of assessment specialists that can be tapped to fill positions on a new assessment staff. It may be possible to raid another staff for a person or two, particularly if salary and potential for advancement or the opportunity to innovate are made attractive enough. Most assessment professional staff will have to be recruited from persons with preparation and experience in educational, or psychological measurement, or allied educational fields. Major competition is with universities, with educational research and development organizations, with intermediate school districts, and with test publishers. National Assessment found that new or near doctoral candidates were a fertile field for recruitment. This is particularly true at a time when new academic positions are few and far between. Good professional staff members are available if a thorough search is made.

Another level of competency that can be utilized for many assessment tasks is the bright young person with a bachelors or masters degree who may not be trained specifically in measurement or research, but who learns quickly. Such persons are particularly plentiful in university and public school communities. *Many assessment tasks can be learned rather quickly by those who are willing to do so.* The key to success in this type of recruiting is a sympathetic senior staff person who is willing to spend a little time in working with junior staff members. National Assessment, again, found this method quite satisfactory in securing some excellent junior staff.

A director of a new assessment program should be prepared to revise his initial estimates of staff needs rather sharply upward if his efforts at program development are successful. The better that he does his job of developing a top-notch assessment, the greater will be the staff needs. If at all possible, several positions that seem to be beyond projected needs should be considered in staff planning.

Consultants. A director of assessment will be in particular need of consultant help in the formative stages of program planning. When considering how best to approach the assessment of science, he may need to have the help of a science educator; when considering how best to secure good questions, he may need the help of a specialist in test construction; and so on. A nearby state university is a logical first source of individual consultant help on a great variety of assessment questions. An intermediate school district or large city school system that employs a variety of professional educational specialists would be other sources of help. Specialists in survey methods or sampling statistics may be sought in psychology, or sociology, or mathematics departments. *One should not limit one's choices to education,* of course.

In addition, there is beginning to develop a cadre of national consultants in assessment who are available for help with new assessment programs. Many of these people are or have been staff members either of the National Assessment program or of a state assessment program. Others have been and are regular consultants to ongoing large scale assessments. Others are staff members of organizations who regularly contract either to help plan or to help carry out large scale assessments.

Contractors. A good source of help in developing an assessment program, either for planning purposes or for development work, is one of the potential contractors of assessment services, such as Research Triangle Institute, The American Institute for Research, Educational Testing Service, Measurement Research Center, California Test Bureau/McGraw-Hill, Harcourt, Brace Jovanovich, etc. Such organizations have many staff members with both training and experience in large scale assessment.

A director of a new assessment program may find that his best and least expensive access to a variety of assessment expertise for planning purposes is through a contract with one or more such organizations. This could very well precede any major efforts at staff recruitment, since the plan to be developed should not be constrained by the existence of a staff with a specific pattern of abilities. Such organizations sometimes rely completely on their own staff for fulfilling a planning contract, sometimes they will seek consultants with special skills.

The National Assessment project can be a source of specific

help for those state directors who want to consider National Assessment techniques and materials for potential adaptation to a state situation. The Utilization/Applications Office of National Assessment offers free consultant service to a state department that wants to examine the National Assessment model closely and for three years has held summer workshops designed specifically for state department personnel working on assessment programs. An under-staffed, under-financed director of state assessment could do much worse than adopting or adapting the National Assessment model, if it provides information of the type desired in a given state. To assume that we need 50 unique assessment programs simply because there are 50 states is like assuming that we need 50 unique math programs or 50 unique career education plans.

Determining Types of Assessment Outcomes

Most large scale assessments have dealt and probably will deal primarily with cognitive outcomes. Certainly the most common starting point for assessment is with either reading or mathematics or both. Not all programs have followed this pattern, however. Some have talked about and a few have implemented assessment in the affective domain, while still further down the list is the area of psychomotor skills.

One of the questions that a general advisory committee should be asked to address itself to is this one—*for what subject area or areas or for what domain should state assessment be designed, and what are the priorities among them?* These decisions are educational decisions, but they are not measurement decisions. These decisions are ones that should be made by a combination of those persons who have the legal responsibility for education (legislators, board of education members, state department personnel, school administrators, and teachers) and those who are served most directly by education (students and parents). These decisions will be based upon feelings of relative importance for assessment of the many areas in which assessment could take place. It may be that an advisory committee will recommend that assessment should be developed simultaneously in 6 or 8 or more different subject areas and in several domains, as was done in Pennsylvania. It may be that an advisory committee will decide to

start with communication skills and mathematics, as in Michigan, or with citizenship and writing, as in Maine, or with science, as in Iowa. It may be that a master plan will be developed, as in National Assessment, to assess in 10 different areas on a 5-year cycle, two areas per year.

In order to make a wise decision on areas to be assessed and on a sequence of assessment, an assessment director will need to provide a variety of information to his advisory committee and to whoever will be making the decisions. One important bit of information is the relative ease of supplying instruments in the cognitive domain as opposed to the affective domain or the psychomotor domain. *Good items and good tests are already in existence in many, but not all, cognitive areas.* Materials in the affective area are much more scarce both in quantity and quality. Materials in the psychomotor domain are even scarcer. Starting an assessment with affective and/or psychomotor domains will put a heavy burden on the staff. A new assessment staff will have enough problems without having to tackle these most difficult assessment domains.

If it is decided to move immediately into some affective area, the policy makers should be made aware of the measurement problems ahead of them. They should be willing to accept developmental costs of materials and perhaps some validation research.

If it is decided to do reading only as an initial assessment (or any other one area or domain), the policy makers should be made aware of the potential psychological impact of such a decision—particularly if there is no long-range plan that calls for other assessment areas of domains. *A state assessment is bound to be in the public eye. Whatever it does will attract attention.* If it is designed as a reading assessment alone, the implication to many educators and laymen will be that reading is the most important educational area and all other areas or domains are clearly less important. This may be exactly what policy makers want to imply. If so, fine; if not, they must realize what the implications of their actions are.

> **Principle:** *The most sensible course of action for determining areas or domains of assessment is as follows: (1) Decide on all areas and domains that are to be assessed in a given state; (2) Develop a multiple year plan (5 or*

10) to cover all of the areas and domains to be assessed; and (3) Select one or two in which to start assessment.

If one wants to maximize the chances of developing a truly quality assessment, it is wise to start relatively modestly and to demonstrate that quality has been achieved before setting out in a variety of directions. The author realizes that political pressures may force a director of assessment to proceed more rapidly than is desired. When it has to be done the director must, of course, try to maintain quality at as high a level as possible. But one should not pretend, either publicly or in private, that any and all demands for assessment-outcome information can be fulfilled at the same level of accuracy and relevance.

Determining Types of Test or Survey Instruments

A major decision that needs to be made early in the planning process is the *type of information that is desired*—information of a normative nature, or information of a criterion-referenced, or objectives-referenced nature. Much has been written in recent years about norm-referenced and criterion-referenced tests. Controversy still exists. The discussion here will not enter that controversy. The author believes that both types of instruments are maximally useful for different purposes, and that a state assessment program should use one or the other type primarily. The worst type of decision to make here is to decide to use a test instrument that happens to be readily available and inexpensive, and then assume that one can use it equally well in either a normative fashion or a criterion-referenced fashion.

A *normative-type test* is maximally useful for providing information about how students rank in relation to each other, or how schools or districts rank in relation to each other, or in any other decision making in which one wants to be able to classify things in order from high to low and to accomplish that ranking in the most reliable fashion. *The standard of "goodness" or "badness" is internal to the test; high scores are good, low scores are bad.*

A criterion-referenced test is maximally useful for providing information about whether students or schools have or have not

attained some objective at a pre-determined level of achievement without regard as to how students may relate to each other in their attainment. *The standard of goodness or badness is external to the test, a human judgment of what is acceptable performance and what is not.*

Norm-referenced testing is analogous to traditional letter-grading of students (A, B, C, D, E) for which one must make decisions between A and B students, between B and C students, and so on. Criterion-referenced testing is analogous to pass-fail type of grading—you pass if you accomplish all minimal requirements, you fail if you don't.

In practice, tests designed to be norm-referenced will overlap with tests designed to be criterion-referenced, but only in part. If one wants to gather assessment information that says, as reliably as possible, that the students in school district A, on the average, are better readers than the students in school district B, on the average, who in turn are better readers than the students in school district C, without regard as to whether any of them can read at any particular minimal level, then norm-referenced tests should be used. If, on the other hand, one wants to find out whether the students in school district A can read at a level that has been judged ahead of time to be appropriate for fourth grades, and the same for students in school district B without regard to whether A surpasses B or vice versa, then a criterion-referenced test is called for.

A state legislator might say, "I want to know which schools in this state are in the lowest ten percent in reading proficiency so that we can appropriate money for reading consultants in those schools" (regardless of whether all or some or none of the schools in the state may already have their students at "acceptable" levels by some judgmental standard). This calls for normative information.

A state legislator might say, "I want to know all of the schools in this state in which students, on the average, are reading below levels that reading specialists agreed upon as being appropriate, so that we can appropriate money for reading consultants in those schools" (regardless of whether it is five percent, or ten percent, or fifteen percent of the schools). This calls for criterion-referenced information.

It is possible to collect both types of information that is

maximally valid by using two separate tests. It is possible to collect both types of information that is somewhat valid by using a single test that is designed to do both jobs. In this author's opinion the latter choice is not a good one, as *it is never wise to plan to do a half-good job.* The best choice is to *design an instrument to do,* as well as is technically possible, *the major, first priority job of information gathering,* and then use that instrument to provide additional information as best it can.

In any event, the alternative approaches should be presented to the policy-decision makers for assessment (CSSO, Advisory Committee, State Board) so that they can decide which type of information they want most. Since this is a controversial area, it probably would be wise to have advocates of both positions present their point of view to the policy makers. Most psychometricians today strongly favor one approach or the other. A director of state assessment—if he is a measurement person—is apt to have a decided preference. He should not let his prejudices dictate a choice by never giving his policy makers the alternatives.

At this point in an assessment program, it would be wise to bring in several knowledgeable consultants to discuss the pros and cons of norm-referenced and criterion-referenced testing for the purposes of assessment for your state. Such consultants will point out a number of practical differences between the two types of tests that should be considered before a final decision is made. They might well point out that there are a number of well-designed norm-referenced achievement tests already available from reputable publishers. In contrast, *there are very few criterion-referenced tests already available.* They will point out if one chooses to use an existing norm-referenced test that copyright laws will call for maintenance of security so that individual items cannot then be released to parents or the press. They will point out that existing instruments rely almost exclusively on the multiple-choice format, whereas criterion-referenced testing requires an open ended format in some cases. They will point out that criterion-referenced materials often are developed to assess basic mastery level of knowledge and skills that all students are expected to attain, and rarely provide information about higher level attainments. They will point out, in detail, how normative test information can be used wisely, and how criterion-referenced test information can be used wisely, and how this relates to local desires.

Principle: *Whether norm-referenced or criterion-referenced type of information is to be gathered in a large scale assessment program should be related directly to basic needs for assessment information.*

Determining Reporting Categories

Reporting categories are those variables on which students and/or schools can be classified as a basis for reporting outcome information. *Some decisions about reporting categories should be made early in the planning process.* These decisions can be reviewed periodically prior to actual reporting of results and some categories can be added or subtracted. However, it is best to have most of these decisions made early as they will and should influence a number of important operational decisions later on.

Grade or age is the first decision that must be made. Does one want information on third graders or fourth graders, on sixth graders or tenth graders? Or would one prefer to report results by age, 9 or 10, 13 or 14, 16 or 17, etc? Most state assessment programs probably will elect to assess (and report) by one or more grade levels. However, National Assessment chose to assess and report by ages (9, 13, 17, 26-35). Maine has followed the National Assessment pattern because they wanted to compare their results to national and regional results. Michigan decided to assess in grades 4 and 7 with two or three other grades to be added in the future. *A choice of which grade or grades in which to assess is, again, primarily a policy choice.* Psychometric concerns would relate to ease or difficulty of data collection at different grade levels, but *the primary concern should be to select the level at which information will be most useful.*

Other potential reporting categories relate to geography. National Assessment reports results by four large geographic areas, Maine by four geographic areas within that state, and North Carolina by three areas. Michigan does no reporting by geographic regions. Some states have rather sharply defined areas that differ either geographically or culturally, by language, or by some other characteristic that may tend to cause people to think of themselves separately. *If policy makers want information about educa-*

tional outcomes separately by region, it should be planned for from the beginning of assessment. Such decisions can only be made within each state.

For years we have reported statistical data—about school buildings and finances and programs—by community size and/or type. For example, we might compare metropolitan areas, urban fringe areas, cities, and urban areas. Or we might compare by community size, by above 200,000; 200,000 to 50,000; 50,000 to 25,000; 25,000 to 10,000; below 10,000. A decision as to whether to report in one of these ways may have political overtones as well as measurement implications. If many people within a state tend to think of themselves as living in one of a limited number of community types, it may be advisable to group schools within these "accepted" community types for reporting purposes. Differences in group test results may or may not appear. And, it may be just as significant educationally to find that differences do not exist between certain groupings as to find that they do.

A common type of reporting category that repeatedly has shown differences in educational outcomes is that of *socioeconomic status* (SES). Research studies have shown that average academic attainment of students from low SES communities is below average academic attainment of students from high SES communities. Most states are heterogeneous enough in SES characteristics that *it would be well to consider,* at least, *some reporting by SES categories.* There will be practical problems of whether to classify by school district or school building or even by individual student. But these can be resolved if it is decided that educational decision making would be improved by reporting such information. Various bits of information can go into an SES classification—such as parents' educational levels, parents' occupations, average income, types of reading material in a home, etc. Collecting such information can be a problem since it gets into the area of invasion of privacy, but even such delicate problems can be handled.

A most common reporting category is sex. But since any scheme of data collection will include both boys and girls, one need not dwell on this decision very long. It is so easy to report separately for boys and girls that it probably should be done if any type of differential reporting is done at all.

Many other categories could be suggested. One could report by

per pupil expenditure (above $1000 per child: $1000 to $750: $750 to $500: etc.): one could report by *pupil-teacher ratio:* one could report by *percent of transiency of student body* or *by percent of racial or ethnic minority:* one could report by *different curricular programs* (new math versus traditional math) or by *open versus traditional classroom.* One could go on and on to discuss various reporting categories, and an advisory committee should do exactly that.

Each separate type of reporting costs money and effort. Each bit of information needs to be secured by one or more questions on a questionnaire. Each additional report requires extra computer time and extra space in printed reports as well as the extra time that staff must give to "program" each classification throughout the entire data collection and dissemination processes.

Deciding which reporting categories to use in a given year is another area in which an advisory committee should be utilized. It would be difficult, probably impossible, for any one person to anticipate the potential outcomes of reporting by various categories. Whether to report results by race or ethnic group is a very sensitive question. People feel very strongly about such questions. Assessment reports should help people improve decision making. It is not their purpose to antagonize people. In fact, antagonism can quickly mask all real utility. An assessment program does live in a public relations atmosphere. *The only sure way to anticipate reactions from the consumers of assessment results is to ask them ahead of time how they would react to various types of reports.* This is not to suggest that we should try to avoid any reporting that may irritate people, but we should consider potential irritation and weigh it against potential good.

Policy makers who suggest or determine a variety of reporting categories should be cautioned that potential differences in results between schools that are categorized in different reporting strata do not, by themselves, establish a causal relationship even though a relationship may be evident. For example, it has been demonstrated repeatedly that student achievement test results from high income area suburban schools are higher than from metropolitan core area schools. This does not meant that if all students from metropolitan core area schools were moved to high income area suburban schools their achievement test results would automatical-

ly go up. *Measured achievement is dependent upon a whole series of input measures* and while one can demonstrate relationship to a variety of factors, changing only one of them will not outweigh all of the other effects.

> **Principle:** *Assessment planning should include careful consideration of a variety of reporting categories. Some modest number of categories should be selected for initial reporting. Other categories can be added from year to year, or alternate reporting categories can be used from year to year on a cyclical basis. Each reporting category selected should relate to the need for information.*

Determining Population

Not only for reporting purposes, but also for data collection purposes it is essential to decide very early in assessment planning the age, grade level or levels for which data will be collected. *The major advantage for a grade level choice is that most schools are organized by grades*, and, therefore, students are already separated by grade and are much easier to reach for data collection purposes. In addition, if the major audience for assessment results is to be the educational community, then grade level results seem appropriate. Educators tend to think of educational concerns by students grouped into grades.

National Assessment chose to accept the more difficult data collection procedures of finding and assessing students by age classifications for two reasons. It was felt that age-grade placements differed sufficiently from state to state so that national statistics would be more meaningful by age than by grade. And also, National Assessment felt that the lay public—the non-educator—could better understand results reported by age than by grade, e.g., that it would be easier for a lay person to think of a nine-year-old than a fourth-grader.

Another point to consider is the increased use of ungraded classes. If this becomes a significant trend, an age designation for assessment could make more sense in the future than a grade designation.

The choice of specific ages or grades for assessment vary from program to program. A "typical" assessment program would select one target group in the upper primary grades, one in the middle school range, and possibly one at the high school level. There seems to be *a tendency to collect assessment data at transition points between educational levels as presently constituted.* These transition points may vary a bit from state to state.

An advisory committee for any large scale assessment program needs to decide at the beginning of a program which *age and/or grade levels* are to be assessed. *This is a policy decision,* not a measurement decision, although reliable measurement is easier at some levels than at others.

Determining Whether Sampling Is To Be Used

The decision whether to test an entire population or use a sample involves a combination of concerns. Clearly there are *policy considerations;* clearly there are *psychometric considerations;* clearly there are *data collection considerations;* and clearly there are *cost considerations.* The best possible staff and consultant thinking on this question should be brought to an advisory committee for them to consider very carefully.

Probably the *most crucial consideration is a policy one,* since psychometrics, data collection, and cost generally would argue on the side of sampling rather than using an entire population. If it is deemed wise for policy reasons to test all students in a population, that preference, typically, will have to be weighed against available resources and technology; so we will consider first the policy implications of the two choices.

One needs to look carefully at the purposes and goals of a specific assessment program in determining whether sampling is appropriate. If all of the specific purposes and objectives of an assessment program can be met by group results, then sampling must be considered very seriously. This is particularly true at the state or school district level.

The only assessment situation that clearly calls for common data collection on all members of the population is when it is deemed essential, for improved decision making, to have exactly the same test information for every pupil in a given grade in a state (or other assessment unit). It is exactly this situation that has

prevailed for years in local school districts that have every-pupil achievement or ability testing at some grade level. Historically, the compulsory state testing programs were examples of this situation, the voluntary programs were not. If a state mandates common testing for all students, they are taking over a role that local districts traditionally have held. This may be good or this may be bad depending on one's point of view of the role of a state department of education. It certainly has important policy implications.

There are many facets to this point, but it should be kept clearly in mind that *it is not necessary to test every pupil at a given grade level on identical material in order to get a good picture of educational outcomes* of groups of students; it is necessary only if one feels that each teacher in an entire state at a given grade level must have the same information for each pupil.

Probably *the greatest advantage of sampling is that for a given amount of effort* (and money) *one can gather more usable information than by using an entire population.* If the goals of an assessment program are to gather statewide information only, it is hard to conceive of any reason for testing all students in a given grade. For example, if there are 50,000 third graders in the state of Limbo, and one wants to gather state statistics only, it is very possible that a sample of 5,000 students (or even 500) would be sufficient if they are selected by a probability sample (to be discussed later). Or, if one can afford to test all 50,000 third graders, and if it is deemed wise to do so, one could select ten 5,000 pupil samples and secure information on ten subject areas, or could go into great depth of information gathering in two or three subject areas. The combinations of possibilities of sampling pupils and content are almost endless.

If one wants district level information, then sampling becomes a different situation. In a school district with one third grade, sampling of pupils is hardly possible for most assessment purposes. In school districts with many third graders, sampling could provide a greater variety of information than common testing on every pupil, in the same fashion as at the state level. Specific decisions of how far to carry sampling should be made only after advice from a sampling statistician. Sampling is a highly developed technical field and the implications of any decisions to sample or not to sample must be reviewed by competent samplers.

Other "compromise" possibilities exist. One could test all students in a population with one short test, while using a sampling approach for other tests. This approach would provide some common information on all students, but would allow for greater depth of data collection over a subject area.

> **Principle:** *Sampling of pupils and/or content should be given very serious consideration for all large scale assessment projects. The only situation when it may not be useful is one in which it is deemed essential to collect common information on all students in a statewide population of students. Sampling should be used to maximize the collection of usable information for stated assessment purposes at the lowest possible cost and effort.*

Determining Data Collection Methodology

Typically, it might be assumed that data collection methodology is the province of an assessment staff; that how to have booklets printed and distributed, that schedules for testing, that administration manuals are operational decisions and not policy decisions. That is true, up to a point. In most instances *it would be wise for the assessment staff,* using whatever consultant help it needs, *to prepare a detailed data collection plan.* At that point it would be wise to have an advisory committee review the plan very carefully. Or it might even be better to constitute a small ad-hoc advisory group specifically for this review process. Such a group should include several school "practitioners" from among those persons who will actually be involved in the data collection process in the schools—building principals, counselors, and teachers. *It is very difficult in a central office situation to anticipate every specific problem that a local school may have* in scheduling testing, receiving and handling test materials, actual test administration, and returning materials to a central location for processing.

It is not uncommon, of course, to contract actual printing of materials, preparation of manuals, and development of data collection procedures to a testing organization that specializes in such

matters. But even in those circumstances, it is wise to have all procedures reviewed before they are finalized. In addition to the fact that potential problems may be identified and eliminated, it also is good public relations to have procedures reviewed by representatives from the schools themselves.

Data collection, like sampling, may take many different patterns. Descriptions will be made of two patterns, a *conventional and an unconventional pattern.* Many variations can exist in addition to the two.

Conventional data collection for large scale assessment *is analogous to data collection for one of the college admissions testing programs.* In a conventional program all questions are printed in test booklets, with multiple questions on each page. A separate answer sheet is used so that results can be machine scored most economically. Manuals are prepared to be self-explanatory. A teacher (or counselor) can read the directions and can understand easily what he is supposed to be doing. Detailed directions are given for: (1) handling of tests, (2) maintaining security, (3) actual directions to be read, and (4) collection and packaging of the materials. Sometimes a school is asked to destroy the tests, sometimes to return them. All communication is handled by mail. All test administration is to groups of students. Group size is related to numbers of students and room size.

These procedures have worked well for specialized testing programs such as admissions or scholarship programs. Whether or not there are any special problems that might be associated with their use in assessment programs is speculative. Perhaps one major difference in the testing atmosphere is that *there* undoubtedly *is less student motivation for taking an assessment test than for taking a scholarship or admissions test.* Whether or not this difference is of sufficient import to call for altered data collection procedures must be decided judgmentally. There is very little research that sheds light on the relative value of differing data collection procedures.

National Assessment's data collection procedures offer a distinctly different model for consideration. National Assessment uses group administration for most of its items, but uses individually administered tests for between one-sixth and one-fourth of its questions. This is done because some of the specific objectives within certain subject areas of National Assessment cannot be

satisfactorily assessed by paper and pencil group administered items. This is particularly true of a series of performance type items, e.g., singing, doing a scientific experiment, producing an art work, and so on. Such data collection is much more expensive than group administered data collection—perhaps ten times as expensive. The choice then must be based upon whether one really wants a particular type of information, because *it is not possible to assess all educational objectives via paper and pencil tests.*

The group administered packages for National Assessment are given to a group of twelve students, which National Assessment's sampling statisticians believe to be statistically optimal. These packages (test booklets) are designed with only one question per page. This was done to attempt to make it look less like a test and to be visually more appealing to students. Multi-colored pictures are occasionally used as stimulus material in some questions. It was felt that these characteristics could contribute to student interest and that interested students will perform closer to their own maximum level of performance. Informal feedback from students has confirmed that many of them do say things like: "This was fun" or "This doesn't seem like other tests." Booklets produced like this cost more money than conventional test booklets.

The group administered packages for National Assessment are administered with taped directions that are played by the persons in charge of administration. This was done for two reasons—to produce uniform reading of directions across the country and to produce a reading of each question and each alternative answer (for multiple-choice questions) without giving any verbal clues. *Standardization of reading the directions is a technique that any assessment program could adopt without excess costs.* The major expense is in the original recording and then duplicating as many tapes as would be needed in the various testing centers or schools. It would require some standardization of tape recorders in various schools. Research has indicated that a national radio or TV type voice was at least as good as (and sometimes better than) the use of a local teacher voice with a local accent. Research also has indicated that a male voice produced better results than a female voice.

National Assessment uses a full time professional staff of

area coordinators for test administration and part-time paid personnel for most of the actual test administration. It does not ask local school personnel to administer its packages. Maine has used a selected group of administrators (SEA personnel) for data collection, and Minnesota plans to use trained exercise administrators. When combined with a sampling approach to data collection, the costs may be manageable, though they certainly are greater than relying on unpaid local teachers or counselors. The major advantage of a staff of administrators who are trained to do the job is the *greater chance of complete uniformity in administration procedures.* An additional advantage is the avoidance of adding one more task to busy teachers and counselors.

While local personnel can do a satisfactory job of administering group tests and following printed directions, an occasional example of maladministration (accidental or deliberate) does occur. Sometimes, a local person wants to be kind to the students by "easing up" on the standard directions, and, in a few instances, a local person may be feeling competitive pressures so severely that he changes the standard conditions deliberately. Research does not tell us whether this happens to a sufficient degree to be of much concern.

> **Principle:** *A director of assessment should carefully weigh standard data collection procedures versus the more precise methodology produced by National Assessment and other projects, and should adopt or adapt as many of the more precise procedures as he can afford and as he feels will contribute substantially to the accuracy of the data. Neither the best possible tests in the world nor the most refined data analyses can compensate for inaccurate data collection. It is essential that the "honesty" of data collection be at such a level that no one can question the validity of the resulting data.*

Determining Data Analysis Strategy

While many of the specific techniques of how to organize and

carry out data analysis can be left to computer technicians and data analysts, there are several aspects of planning for data analyses that need to be considered. The most important of these early decisions is to *determine the specific requirements of the major reporting categories that are anticipated.* While reporting strategies can be altered throughout an assessment program, and even after the data collection is complete, some options may be missed if they are not built into the early planning. For example, if one is using a sampling approach and wants to report precise results for some very sparsely settled area of a state, it may be necessary to oversample students in that area. This would need to be planned for in the sample design, well before data collection takes place. Or, if one wants to report results separately for students from high SES (socio-economic status) home or areas, from average, and from low SES homes, one must plan to collect this information either from students, school records, or principals.

There are an almost *infinite number of variables that can be used to categorize students and/or schools for data analysis.* A director of assessment will need to consider very carefully what the major audiences are and then relate the decision making about data analysis categories to them. For Minnesota assessment, the advisory groups were polled—category by category—to determine their feelings as to the relative importance of a great variety of potential reporting categories.

Such decisions are *clearly policy decisions.* Psychometric concerns are primarily those of the collection of the information necessary to do the requested categorization. Good SES information, in particular, can be very difficult to collect if it must be done for each student. But if school level information is adequate, the data collection is greatly simplified and probably is much more accurate. Problems of invasion of privacy are raised if certain questions are asked directly of students.

A director of assessment would be wise to secure consultant help from persons who have worked extensively with data and who have a knowledge of its real utility versus its "hoped for" utility. It often is difficult for a person who is asked to offer advice on the importance of various reporting categories to see all of the potential advantages or disadvantages. Most typically, people who do not deal with categorized information regularly agree that almost any reporting strategy that is suggested is good—on the basis that it

"might be nice to know," or on the basis of strictly personal interest. Realistic consultant advice, then, should go directly to an advisory group and should become a part of their own decision making.

In addition, whatever plan is developed for data analysis categories should be reviewed carefully by several competent data analysts before it is finally adopted. Most assessment specialists know quite a bit about data analysis, but only a few are really data analysts. The inputs of a good data analyst are important in every phase of assessment; they are crucial in this phase.

> **Principle:** *Advisory groups should determine which reporting categories are essential, highly valuable, valuable, or not necessary, in consultation with the director of assessment, the staff, and consultants who provide the necessary information about data collection and data analysis problems and costs.*

Sometimes in the early planning phase the director of assessment, together with the staff, will need to decide on and *arrange for a computer system for actual data processing.* If a state department already has its own hardware or is already tied into a nearby system, the director of assessment may have to fit the program into that system. This is good if the system is good, and bad if the system is not efficient for assessment purposes. If a system does not exist to which assessment can be tied, the director will have to investigate available systems and select the one that gives promise of being best for use.

In either situation the director will need special consultant help. *Data processing can be very irritating if it is not designed to process assessment results speedily, accurately, and in a way that provides easily usable printouts for the assessment staff.* While it is impossible to anticipate all data processing problems prior to actual implementation of a system, as many "bugs" as possible should be worked out of the system ahead of time. The system should be developed and tried out while all other assessment procedures are being carried out; it should not wait until data are ready to be analyzed. A good computer specialist should review all data collection and data analysis procedures prior to their adop-

tion, in order to look for potential problems that may come up later, after one cannot alter the procedures. For example, the process of getting information from answer sheets through a document reader into a computer should be reviewed early enough in the planning process so that answer sheets can be designed to maximize efficiency of the total process.

Determining General Reporting Policies

In the rush to develop a sound assessment plan, a director is very apt to concentrate so much on developing instrumentation and on data collection that he does not pay enough attention to the end product of any assessment—the resulting information that needs to be widely disseminated. *The ultimate success or failure of any assessment depends upon the utility of the information that it generates.* Utility must be planned for, just as much as writing a test or administering a test.

The general advisory committee for assessment should be particularly useful for purposes of developing reporting policies. Such a committee will certainly include persons who represent most of the potential users of assessment results. These are the persons best suited to tell a director not only what information they want, but, also, how they want it reported.

Any large scale assessment undoubtedly will report most of its basic results in tabular or list form in one or more tome-like volumes. This is essential, so that all basic information is available to those persons who want and need it and can read tables of results easily. But for typical parent or teacher or administrator or board member or legislator this is not sufficient. Shorter, more focused reports and summaries will be needed.

An advisory committee cannot develop the details of a reporting plan, but it can provide advice as to the major audiences that need to be addressed, the timing of reports, and the specific type of content to be reported.

Even in the earliest phases of assessment planning, the director should produce or have produced a few sample reports, based on hypothetical results from the assessment. Such samples help to focus the attention of an advisory committee away from the generalities of saying that "We want to know if our students can read," or "We want to know how achievement varies from high

SES schools or low SES schools," and directly into what can be produced from assessment. One source for such samples could be the existing reports of state assessments and of the National Assessment project.

Bridging the gap between the desires for assessment information expressed in generalities and specific examples of tables of results can be a difficult one. In order to avoid publishing information in a form that no one wants or can easily comprehend, the best approach seems to be to focus directly and specifically on sample results and then ask the advisory group, "Is this what you want?", "If you receive information like this, can you use it, (or can others use it) to improve education?"

> **Principle:** *It would be highly desirable for a team of communication specialists and measurement specialists to produce a few, short, mock reports of assessment results during the process of planning the entire program. These reports could serve as a resource in determining the final form of the program and could be used to explain to people what assessment is all about.*

In this area of assessment, as in other areas, there are specialists: persons who know best how to provide information in forms that will communicate most effectively. Many large school systems have one or more staff members whose job it is to feed information about the schools to the supporting community. An effective large scale assessment program needs the same type of staff. Neither psychometricians nor statisticians are noted for their expertise in communicating results of their investigations to non-technicians. During planning procedures a director of assessment generally will be able to secure what help is needed from consultant sources, but before data starts coming in, one or more staff with specific expertise in communicating assessment results to its various audiences will be needed.

Determining Costs and Alternatives

No director of assessment is going to have a free hand to plan

an assessment without the constraints of budget restrictions. And almost invariably, a group of advisers will ask for a program that would cost considerably more than available resources. It is easy and logical to say "yes, that would be good information to have" and to say it repeatedly, without thinking too much about the dollars that must back up every such statement.

Sometimes a director of assessment is told that he will have $100,000, or $500,000, or $1,000,000 to use to develop a program. Such a situation is relatively easy to manage because activities can be tailored to a set maximum, and then a plan presented that can be bought for $100,000 or $500,000 or $1,000,000. The major problem with such an approach is that most advisers will want more information than the plan calls for. But that, too, can be dealt with rather easily, from an administrative point of view. A director can either leave the program as originally designed, or can make substitutions in the plan by eliminating something in the program whenever adding something, or he can secure more money for the additional things that are added to the plan. This situation is not too bad for a director of assessment because he has the support of his advisers if the decision is made to seek more funds than originally allocated. Such an approach may offend those who feel that the starting point should always be what is best for education rather than "how much good information can I afford to buy?" However, we make all other *financial decisions in education in terms of what we can afford,* so there is considerable logical justification for approaching assessment the same way.

A major danger with this approach is that one may be tempted to *reduce the quality of instrumentation, data collection, or analysis* in order to get "more" into the program. *This should be resisted with all possible vigor.* While one cannot determine some absolute level of quality of information to be obtained, one certainly can set standards below which it is unwise to go and still do an assessment. If it is not possible to buy—for $100,000—reliable information that decision makers in a given state desire, then *it is better to abandon assessment than to buy unreliable information* for $100,000.

The alternative approach to determine costs is to *develop an assessment plan based on obtaining all of the information deemed essential by the decision makers, to build acceptable quality levels*

into the program, and then to put a price tag on it. When this approach is used, the chances are very great that the price tag will be far above anything that a CSSO may have had in mind. The potential advantage of this approach is that the initial internal estimates of cost may be raised significantly, and one may end up with more money for assessment than if one arbitrary figure is chosen originally. A serious danger with this approach is that expectations for quantity of results may become so firmly established that quality of instrumentation or data collection may be lessened in order to save money, rather than reducing the scope of the program.

Whether to try the first approach or the second or something in between probably will be dictated by local conditions and by the way that one must operate in a given situation to maximize an allocation of funds while maintaining quality of the program. *One must operate within the constraints of local political concerns.*

The author tends to favor a variation on the first approach and to start with a relatively firm potential budget for assessment. To obtain a realistic initial budget, one can profit by the experience of other states that already are doing assessment. If one finds that the average annual budgets of ongoing state assessment programs are $500,000, it is possible to approach a CSSO with that information and to ask for some reasonable amount as a planning budget. A large state might start higher, a small state might start lower.

With a given figure in mind, a director of assessment then should develop a plan for obtaining quality information within that budget, without too much regard to quantity. But in addition he should plan one or two alternative budgets that provide for additional information that his advisers desire. This, then, provides the CSSO with several choices as to what can be obtained for different dollar amounts.

With this approach, one may run the risk of starting assessment with a smaller program than most advisers really want. But this may be the best thing that could happen to assessment, if it does result in a program that produces usable, reliable information that is used by decision makers, no matter how limited that information may be. Eventually an assessment program will stand or fall on its real utility, not its potential utility. If one can get a small amount of truly good information for $100,000, it is easier

to justify $350,000 the next time around than if one starts a program by obtaining large amounts of questionable information for $150,000 or $200,000.

In preparing initial budget estimates, a new director of assessment should very seriously consider seeking consultant help from persons or organizations that have already been operating or participating in large scale assessments. It is very difficult to anticipate all of the expenses associated with assessment unless one has been through a program. Spending a few thousand dollars at this phase of the planning could save tens of thousands later and could save a multitude of headaches associated with poor projections of costs.

Determining Time Schedule

Until one has been through the multitude of details of running an assessment program it is very difficult to establish a realistic schedule of activities. The closer the program conforms to a traditional large scale testing program, the easier this task is. But the more one tailors an assessment program to specific state needs, and the more one plans to develop new materials or methodology, the more difficult it is to establish a firm schedule. Consultant help from persons in other states or from organizations that contract for large scale assessment can be very helpful.

The major advice that can be given is to produce a flexible schedule and to allow for more time to carry out the program than original estimates suggest. If the director buys or develops a relatively conventional program, schedules can be developed rather realistically ahead of time. But if a person plans for a program that requires developmental work, he will be at the mercy of the individual schedules of a great variety of consultants. For example, the mere problem of finding an appropriate date for getting five or six consultants together may result in a master schedule being slipped by a week or two or even a month. If the director relies entirely on staff for decision making, it is easier to hold to a predetermined schedule, but the potential sacrifice of quality by never using consultant help to generate useful ideas is much too great to allow a schedule to dictate program decisions.

Chapter 4

DEVELOPMENT TASKS IN ASSESSMENT

It should be possible to point up and discuss an almost limitless number of specific details of actually developing and running a large scale assessment project, even after all of the vital planning decisions are made. But even in such a detailed discussion, it would not be possible to anticipate every specific question that would arise. Therefore, this section will be limited to major operational decisions that will be faced in almost any large scale assessment.

Developing Assessment Objectives

There are several things to consider in developing specific purposes, goals, and objectives for whatever area or areas or domains that are going to be assessed. The major general questions are, (1) *who will develop the objectives*, and (2) *what process will they use?*

The two questions involve both management of the developmental process and the development work itself. *Management can be handled in-house or can be contracted out; developmental work can be handled in-house or can be contracted out.*

If a director of assessment decides to manage the development of objectives *in-house*, some person must be assigned time to do the job or be hired to do the job. This task does not require a subject matter specialist; it does not require a measurement specialist, though either type may be used. It *requires an organizer* who knows enough measurement to avoid the pitfall of developing completely unmeasurable, "pie-in-the-sky" objectives. It requires someone who works well with curriculum and subject matter specialists. Editorial competence is most advantageous. Given sufficient time, such a person can handle the managerial aspects of developing objectives in several different areas.

If a director decides to go *outside* for this management task he will need to develop a request for proposal (RFP) that spells out in some detail the desired outcomes and probably the process to be used, although the latter can vary a bit in amount of detail. There are a number of organizations that specialize in pre-

paring objectives, so resources are numerous. This approach (which has been used extensively by National Assessment) makes fewer demands on an in-house staff, and is particularly useful for a beginning assessment program. It must not be assumed, however, that this approach eliminates the *need for a staff person to monitor the contractor's activities and his product.*

Assuming that *management is in-house*, the process used to develop the objectives can vary considerably. A state department could elect to use its own staff of curriculum consultants (assuming that such a staff exists) to prepare initial drafts of objectives. To some this has the advantage of being a non-budget item (on an assessment budget). But it is a budget item in that a state department staff member cannot be working on assessment objectives while still carrying on other activities at full speed. If competent people are available on staff and are willing, this is not a bad approach. But if the task is looked upon as an overload that has to be sandwiched in somehow, quality of the product probably will suffer. Another approach is to *form small committees of professionals in a given subject area*—from schools and universities and other good sources—*to draw up a set of objectives.* If one can secure the cooperation of a statewide professional organization, such as the mathematics educators or reading specialists, considerable expertize can be focused on the task. If one can get voluntary assistance, this is very easy on the budget. In many instances one must be prepared to hire consultants for this task.

If it is decided to go to an *outside contractor for management* of objectives-development, there still are two process options. One may allow the contractor to use his own staff to develop the objectives, or one may require that the contractor use professionals outside his staff. Good statements of objectives have been developed both by in-house staff (of an SEA or contractor) and by part-time consultants. It seems much more crucial that people be competent than that they be staff members of some given organization.

A state department may deem it politically expedient to have objectives developed by persons inside a given state. Such a process certainly makes the potential acceptance of the resulting objectives more palatable to some persons than if they are developed elsewhere. Other things being equal, this certainly is an appropriate condition.

Developers of objectives, whoever they may be, may start from "scratch" or they may start by reviewing presently existing statements of objectives. *There are so many general statements of subject matter objectives that have been developed in recent years, it hardly seems worthwhile for each of our fifty states to repeat the process in toto.* In Maine, the National Assessment objectives for citizenship and writing were reviewed and adopted as they existed. In Connecticut, the National Assessment reading objectives were compared to the state's already established objectives. For areas of overlap a '71-'72 Reading Assessment was developed that used selected National Assessment reading exercises. In Michigan, performance objectives in mathematics were developed by a statewide committee of mathematics educators, under the auspices of the state department.

The Instructional Objectives Exchange (IOX) at UCLA is one source of objectives; the Westinghouse Learning Corporation (through the Measurement Research Center) is another source; National Assessment is another; various state departments, such as New York, already have stated objectives; and some individual school districts have well developed statements of objectives. Certainly *one should look at existing statements and have them reviewed carefully before launching into full scale development*—unless the process of development is deemed essential in a given state. Some people feel that developing one's own objectives has a distinct advantage because it forces people to think carefully about their own purposes.

After statements of objectives have been developed by whatever persons or processes selected, it is highly desirable that they be reviewed carefully by persons with a variety of backgrounds. Subject matter objectives are not (should not be) the sole property of subject matter specialists. While a fourth grade teacher is vitally concerned with the reasons for teaching reading and what is being taught, so are the students and their parents, the principal, the superintendent, the school reading consultants, members of the board of education, taxpayers and state legislators, etc. *Why we teach what we teach is of vital concern to anyone and everyone interested in education* and interested in the school as one agency for promoting education. If this position is accepted, then *all assessment objectives should be reviewed by a variety of persons who represent both professional educators and laymen.*

National Assessment found reviews of its objectives (in face-to-face situations as opposed to mail reviews) to be a most valuable addition to the development process. Professional educators always had suggestions for improvement; lay persons sometimes did. However, lay persons always felt it was most important that they be involved in the review process. Lay persons were particularly sensitive to objectives in social studies, citizenship, art, and music and were least sensitive to objectives in mathematics and science—sensitive in the sense of making substantive suggestions for improvement. Of particular note in National Assessment is the *use of students in the process of reviewing objectives*—generally high school students or college undergraduates.

Above and beyond the values of improving objectives by reviews by various people, the public relations aspects of an assessment program can be enhanced by being able to say that objectives have been approved by parents and board members as well as by teachers and curriculum specialists. *A review process is time consuming* and, thus, causes scheduling problems. It can be expensive if one pays reviewers, inexpensive if their work is voluntary.

It is important to have objectives reviewed whether they are adopted or adapted, or created. Ultimately, a variety of people should say, "Yes this is what we want our children to learn, know, and be able to do and feel."

> **Principle:** *Production of well developed, succinct, easily understood objectives requires a lot of work, particularly if they are developed completely from "scratch."*

Using Existing Instruments

As soon as objectives are developed, the next step is to move to *instrumentation*—to secure the test, or tests, or other data gathering devices that are to be used in assessment. One approach is to look for existing instruments that have already been used or can be used in assessment.

In Iowa over 90 percent of the schools administer the Iowa Tests of Basic Skills (ITBS) in one or more grades as a part of the voluntary statewide testing program. This has been in operation

for many years. If it is decided that one or more specific assessment objectives can be met by results from the ITBS in Iowa, then the data are already collected and it would be possible to move directly to analysis and reporting. If one objective is to rank order schools on reading, or arithmetic, or study skills, or language skills, then it could be done in Iowa with minimum effort. If it were desired to identify districts where there are a high percentage of low achieving students in one of these subject areas, then the data are available.

Most states, however, do not have results from a common test already being given to most students. In one specific area there is the possibility for securing statewide information, however. *The "Anchor Test" projects being funded by the U.S. Office of Education and carried out by Educational Testing Service have provided a link between a variety of standardized reading tests at grades four, five, and six.* Norms tables have been prepared that provide for estimating a score on the anchor test, regardless of the reading test actually used. Thus, for summary statewide information in reading it might be possible to use the results of existing standardized reading tests at those grades and to extract some useful results.

A more viable option for use of existing instruments is to select one or more readily available standardized achievement tests or portions thereof and then administer those portions to a sample of students. If the major purpose of assessment is to gather rank-order type information for pupils or districts, then a standardized test that maximizes differences between results and has content validity for one's assessment objectives is a realistic, and inexpensive choice. *The key to this decision is that the test selected must be considered to be a good measure of one's objectives, and a norm-referenced test must be considered appropriate.* Pennsylvania chose to use existing tests for some of its goals and chose to develop material for some others.

If one prefers to describe attainment to a greater extent than to compare it across pupils or districts, there also are some choices of existing instruments. National Assessment items have been used in assessment programs in Iowa, Maine, and Connecticut, for example. Several commercial publishers have new, criterion-referenced items and tests on the market this year. These tests or items must pass the same "test" of content validity for one's own state assessment before adoption. Such tests will be less efficient for comparative purposes across students or districts than stand-

ardized tests, but will be more efficient for judging whether some judgmental standard of attainment has been reached—whether most students have "mastered" arithmetic skills or basic knowledge of science.

For years the only available achievement test options available to evaluators have been standardized norm-referenced tests. We now have several major publishers who are marketing criterion-referenced tests, and, in addition, National Assessment's released items are available free of charge to anyone caring to use them.

Developing Instruments

Test development is a well established procedure. If a director of assessment and his advisers decide that existing instruments would not meet their needs for assessment information, then test development is called for either in whole or in part. Such a step should not be undertaken lightly because *good test development is a time consuming, costly process that will result in quality instruments only if carried out in a professional manner.* Hurried production of test questions and inadequate review of items can result in an inferior instrument that will produce inferior information.

If it is decided to go the developmental route, *the easiest approach is to collect existing questions from one or more sources and then put them together as a test.* This process requires that each item examined must be judged to be a good measure of one of the specific assessment objectives and its known or assumed item difficulty must be judged to be appropriate for assessment needs. This is *essentially the same process as selecting objectives from a large list of objectives.* An advantage to this approach is that the items being evaluated already have passed several screening checks and, in all probability, are basically sound items that may or may not be appropriate for specific purposes. If such items seem good enough and if item statistics are available, they may be accepted for an assessment test without further investigation. If they are at all experimental or untried items, they should be put through a formal tryout before adoption.

This selection process probably is best done by a combination of subject matter specialists and professional item development personnel. The author has found that a working committee of two

or three subject matter specialists, plus one or two measurement specialists, can form a well-balanced group for evaluation of test items one by one.

There are a variety of sources for such items. The entire pool of released National Assessment items is available without charge to anyone. These items already have national p-values* and have been reviewed by from twelve to twenty professional reviewers. Test publishers have files of items that they are willing to sell or rent for state assessment purposes. The IOX at UCLA has published many sample items that could be used. Various professional organizations have produced tests and sets of items that could be used as a source. Some state departments have already developed tests for their own uses and might be willing to loan, rent, or sell items to other state departments.

A major point to be considered in this borrowing or buying procedure is the *locus of legal ownership.* National Assessment has not copyrighted its items so they may be used at will. But publishers do copyright their material. A director of assessment must decide whether it is important for the project to be the "owner" of its test items and whether lack of ownership will jeopardize security of the items. *It is essential to be confident that items to be used in one's own assessment will not have been administered to exactly the same students before—at least not recently.* Unless National Assessment items begin to be reproduced widely in sources that teachers read regularly, it probably is not necessary to worry too much about those items. But test items from an IOX booklet that may have been widely disseminated to all social studies teachers in a state might not be appropriate. We sometimes fear the loss of item security too much. Such fears are most justified in testing situations that are competitive in nature, less justified in other situations.

Freedom to use one's items (and tests) in whatever fashion is deemed essential to success of the assessment program is very important. If education is enhanced by publishing and distributing test items with the assessment results, then a director should be free to do so. Items which cannot be used freely to further effective reporting of one's own assessment rarely should be considered—only if nothing else is available that will measure the

*percentage of pupils responding correctly

objectives to be assessed.

If files of items with high content validity for one's own assessment are not readily available and usable, a director of assessment will be forced into an *item development process*. Such a process *should be undertaken only after deciding that available sources are not satisfactory*. There is no psychometric reason why each of the fifty states needs to create a set of reading items, and a set of mathematics items, and a set of social studies items, etc. While each state may think of its educational goals and objectives as being unique, that is a highly questionable assumption.

If development of items is to be done, there must be people, time, money, and other resources to do it. A staff person who works with development of objectives is a logical "manager" for test development also. Such a person should have some knowledge and skill in the psychometrics of test development, in order to be able to see that appropriate quality checks are made. This person does not have to be an item writer or a professional researcher. Rarely would an assessment staff within a state assessment project be so large that one would seriously consider including item writers, except perhaps on a part-time or limited time basis. It might be appropriate to hire a good item writer in social studies for a few months, or a number of item writers for a week each, distributed over several months. The latter approach is more typical. Test publishers often contract for test items on an item by item basis, and pay a set dollar figure for each one.

A staff item development manager(s) is essential; staff item writers are not essential. In securing specific items, it is possible to go directly to some test publisher or other organization that specializes in producing items, or to go directly to potential item writers. If one goes through a contractor, one is buying existing expertise in item development and must pay for that expertise. It is not appropriate, however, to assume that such costs can be saved by going directly to item writers and carrying on all item development procedures directly. To do it directly requires all of the developmental time built into one's own staff and that is effective only if one can secure staff that are as competent in test development as a potential contractor's staff. *A beginning assessment program should very seriously consider going the contractor route*, until and unless staff expertise is equal to a contractor.

A major concern with completely internal development of

test items is lack of objectivity in judging their worth. A developer of test materials cannot be considered an appropriate evaluator of those materials. It is not easy for some people to see their test "progeny" torn apart by critical reviewers.

Whether handled internally or externally, *item writing is better done by a number of persons than by a single item writer.* National Assessment found, by bitter experience, that a single item writer for a subject area often produces a set of items that reflects his own educational biases. While such biases can be weeded out by an appropriate review process, it is better not to let it happen initially.

In the author's opinion the best item writers, by and large, are subject matter specialists with some psychometric "know how" at a practical level. Such persons can mentally scan the entire subject field for ideas and can also avoid the most gross type of writing errors that often are made by persons without psychometric experience. Such an approach, of course, must be buttressed by careful reviews by professional psychometricians as well as by other, equally competent subject matter specialists.

It is almost impossible to prejudge whether a particular person will be a good item writer. One approach to the selection of item writers is to ask ten or fifteen subject matter specialists to prepare three or four or five items each and review the items to decide which persons are producing the best ones.

Development of each and every test item should include specific scoring specifications. For a multiple-choice item this is simple. For an open-ended item it may be simple or it may be very complex. In either case *it is essential that item development include the production of the answer, or answers, or samples of answers that are acceptable and those that are not.* This must not be delayed until after data are collected. *One should never accept an item for use in assessment without a complete scoring rationale.* Occasionally an unscorable item is developed; occasionally the only feasible scoring procedure is so complex or so expensive that the item may not be worth retaining.

There are a multitude of specific do's and don't's in test development. The reader is referred to Finley and Berdie's *The National Assessment Approach to Exercise Development* as a reference on item development in National Assessment. And if he becomes serious about handling item development in-house in a large scale assessment project, it would behoove a director of

assessment to seek access to National Assessment's staff manuals on item development. A director of assessment who must have new items created, either by contractor or in-house, must develop a similarly detailed manual of test development procedures to insure production of items in a manner that will be acceptable to him. Consultant help may well be needed for this activity.

Regardless of who produces *items for assessment*, they *must be reviewed and tried out and reviewed again before being accepted for use*. Initial reviews should come after items satisfy the author and after they satisfy a first scrutiny by staff (contractor or in-house). These reviews should be by subject matter specialists primarily, but also by psychometricians. One looks first for content validity—is this item measuring some specific skill or information that represents direct attainment of a given objective? If the answer is no, then the item should be shelved; if the answer is yes, then the item should be considered further as to whether it meets other *criteria of acceptability*. As part of the judgment of validity, reviewers need to be concerned with potential item difficulty. A very difficult item might be judged unacceptable for that reason alone, if one doesn't want to assess difficult or complex concepts.

Other criteria for judgmental acceptability relate primarily to *understandability* and *extraneous cues*. It is essential to have items that are worded so clearly that the very poorest reader can understand the question or task. *The choice of simple vocabulary and simple phraseology is essential in order to be confident that resulting answers do adequately represent a student's "true" level of knowledge or skill.* It is essential to have items devoid of extraneous cues that could lead a student toward the answer, even if he doesn't know it. And it is essential to have items with response modes that are simple enough for all students to handle.

Knowledgeable adults, who work with students regularly or who work with test items regularly, can do a fairly good job of reviewing and editing items for their understandability. But no one can anticipate every potential problem or every potential cue. *There is some evidence that one can come a little closer to the ideal by including students in a review process.* They sometimes see problems with an item that adults do not.

Still another point of concern is *item sensitivity* and *potential feelings of invasion of privacy*. Michigan received some unfavorable publicity in relation to a reading selection that was interpreted by some persons as presenting big business in an unfavorable light,

and another that was seen as being divisive, calling attention to economic differences between groups. National Assessment attempts to meet the problem of sensitivity by having all potentially sensitive items reviewed by panels of lay persons—who were asked specifically to react to item sensitivity.

But human judgments cannot do the job completely. Following initial reviews, items should be tried out with small, representative samples of students of the same age or grade for which the items will be used. The major purpose of tryouts is to:

- Look for ambiguities that haven't cropped up before;
- Look at some simple statistical relations in search of problems; and
- Seek some rough estimate of timing.

If items are being developed to be used in a normative fashion, particular emphasis is put on the statistical characteristics of the items. If items are being developed to be used in an objectives-based fashion, particular emphasis is put on identification of ambiguities. These are not exclusive goals, of course.

For a norm-referenced instrument, items are needed that generally are in the middle ranges of difficulty (p-values between .40 and .60 or between .35 and .65). For a criterion-referenced instrument, items are needed that are close to the level of assessment decision making, e.g., *if it is important to determine what has been mastered in a given area, items are needed that are answered correctly by most students.*

For a norm-referenced selective instrument, items should discriminate between high scoring and low scoring students, and items that do not are rejected. For a criterion-referenced instrument, it is important to take a careful look at items that do not discriminate, for evidence of ambiguities, but they are not rejected if they still pass the judgmental tests of content validity and understandability.

For a criterion-referenced situation, it might be wise to interview individually some low-scoring students to see whether they could verbalize why they answered items as they did (to see if such answers were based on extraneous cues or misunderstandings of directions or specific words, etc.).

For an assessment test that is to differentiate between before-instruction and after-instruction situations, one would seek items that worked that way statistically in tryouts. Most of the present instruments that are used for pre- and post-testing have not been developed that way. The best that can be hoped for from such tests is evidence of some general increment of learning, rather than learning related to specific objectives for a given year or given term.

The results of tryouts of items should go to a small subset of the reviewers who did the first general review. This procedure builds upon the familiarity with the items that they already have acquired. In some instances, staff might screen out the items that definitely and obviously worked and only take the difficult cases to the reviewers. Clearly this process has to stop some place. In the author's opinion, the process should stop with reviewers, not with a statistical analysis.

> **Principle:** *No matter who writes test items, they must be reviewed and tried out and reviewed in order to assure maximum utility.*

After items are finally selected for use in an assessment program they must be packaged—put together into a test or tests. This procedure should be a joint one between an assessment staff member and someone from the scoring agency. This statement assumes that machine scoring will be utilized for choice-type items and that, if open-ended questions are used, the same agency will be responsible for scoring all of the items. It is very common to secure scoring services from a contractor, though it doesn't have to be that way. If a state department already has a facility for machine scoring of test answer sheets, then assessment could tie into it. This will be discussed later. The point here is that the tests themselves should be designed to be attractive, to present items in a psychometrically satisfactory manner, and to maximize the ease of scoring.

A major point to be considered *in the packaging process is to avoid putting two items in the same package (test) if one of them provides a cue to the answer of the other one.* Another point is to *arrange items so that directions for administration can be simple and direct,* by grouping all items that require the same response mode. A traditional pattern or arrangement of items calls for

moving from the easiest to the most difficult. But National Assessment found that a random arrangement of items by difficulty level was superior for its purposes. In a random arrangement, a difficult item is apt to be followed by an easier one. This seems to maintain motivation better than a situation in which items become increasingly more difficult. Most tests tend to crowd items rather close together on a page. National Assessment went to the opposite extreme of putting only one item per page. Paper and printing costs are not the major expenses of an assessment program, so at least *some attention should be paid to the esthetics of item spacing, type size, and illustrations, etc.*

This process of item packaging is a staff plus scorer task. It requires persons who work very precisely and accurately. If well done, it contributes significantly to good data collection; if poorly done, it may result in less valid data collection.

Selecting a Sample

If it is decided to use a sampling approach to data collection rather than an every-pupil approach, there are a few major decisions about specific details. A general publication on assessment cannot attempt to cover the myriad details of sampling; it only can attempt to cover the major aspects of sampling that a director of assessment will face.

The major contribution that a director of assessment can make to a sampling plan for his state is to *insist that it be a probability sample rather than some convenient, but non-probability sample.* For most non-statisticians, this is comparable to saying that one must have a random sample. Although random and probability samples are not synonymous, the "intent" of both are the same. If a director of assessment tells a sampling statistician that he wants a random sample, the statistician will understand his intent and will help to design a sample that does the job that is desired. The intent of a probability (random) sample is to maximize the accuracy of results and to allow a person to develop and speak of the level or degree of error of statistical precision that may be present in the results.

To maximize accuracy and statistical precision is a noble goal, but if that is the goal, why wouldn't it be better simply to test all students in a defined population, rather than test only a sample, and thus have zero error? There are at least two reasons for con-

sidering the use of a sample rather than testing an entire population. One of these is the fact that the *costs and effort to test some populations may be so great that it is better to accept some error than to pay the price of total testing.* National Assessment's sample for in-school students is about one student per thousand for any given item. This results in a standard error of less than five percentage points in item statistics, such as, "78 percent of all students can answer the following question (followed by the specific question)." To test all students of a given age in the country would be impossible, (practically), and even if it were done could cost (hypothetically) as much as a thousand times more in administration costs.

And even then it probably would not eliminate all error. Different types of error may be present or may be exaggerated in every-pupil testing: (1) the errors of administration; (2) of not following directions exactly; (3) of not following timing exactly; and (4) even of cheating. Such *errors are almost certain to be greater in any mass testing than in smaller, well-controlled sample testing.* A very well-designed small sample and very precisely-executed data collection may, in fact, produce less error than a mass, relatively uncontrolled testing program. Probably every reader of this publication has first-hand knowledge of instances of mal-administration in large scale testing. This is the second reason why sampling may be better than every-pupil testing.

If a director of assessment insists on a probability sample, he will be eliminating certain types of methodology for doing a sample. He will eliminate the use of a voluntary sample—those school districts, school buildings, or students who volunteer to participate in assessment. For a voluntary sample he could never compute any meaningful sampling errors, and, thus, the results could never be generalized beyond the districts or buildings or pupils that were tested. In like fashion, an administrator-selected sample yields no meaningful inferences beyond the schools selected. If a school district contains 25 elementary school buildings it is not appropriate to select—by personal choice of the superintendent or one of his assistants—the 5 schools to represent that district in a statewide sample. Neither is a convenient sample appropriate—those schools within fifty miles of the state capitol. In fact *no sample that relies in part or in whole upon human choice* of which students or which districts will be in a sample *is a probability sample*, and thus *no such sample would be appropriate.*

One of the first questions that a sampler will ask of a director of assessment is the reporting categories that he wants to utilize for his reports of results. While this question may not seem to be a sampling question, it does have implications for sampling. If a director wants to report results for all schools in one region of the state where the population is low, it may be wise to stratify schools by geography and then draw a sample from each stratum. If a director wants to report results for schools with varying pupil expenditures, it may be wise to stratify schools by pupil expenditures and then sample within each stratum.

> **Principle:** *Stratification of a sample is important to do if there is any concern that some pupils or schools from some location, or with some other characteristic might not be adequately sampled by a simple statewide sample. Sometimes a state with a single large major city may need to define that city as a separate stratum to ensure its adequate representation.*

Some reporting categories need not be of any concern for sampling purposes. Sex is an example. The number of boys and girls in schools are so close that any good probability sample will include adequate numbers of both. *It is not necessary to stratify a sample by every reporting category that one desires, but it may be necessary to stratify by one or by a few.* A good sampling statistician can make that decision after he knows the categories and also has adequate information about population statistics for the desired categories. *Stratifying is just one technique that can be used to maximize the efficiency of a sample.* A sample is most efficient when it is designed to produce a desired level of statistical precision for the least cost in both money and effort.

A technique for producing unbiased estimates is a *weighting technique.* Suppose that a state has 50,000 fourth graders, of which 5,000 are American Indians. And suppose further that it is deemed important to report results separately for American Indians. And suppose further that a 5 percent statewide sample (2500 students) would produce satisfactory statistics for all other reporting categories. But 5 percent of 5000 is only 250 students and this will not produce the desired level of statistical precision for the Indian population. In order to produce more precise

results for the American Indian sub-population, a sampler could double the sample of Indians, for example, to 500 (10 percent of the Indian population). But it isn't necessary to double the total sample size (from 2500 to 5000) as that would double the costs of data collection. Statistically the sampler can simply enter the 500 Indian student sample results into the total estimates at half the weight, 10 instead of 20,* in order to maintain the correct total ratio of Indians to non-Indians.

In fact, *in a real probability sample, each and every individual student will have a "weight" associated with his own test results.* That weight is his individual probability of actually being in the sample that is drawn. His actual test results go into the total sample results in proportion to his individual weight. It is this procedure that produces a weighted probability sample rather than a random sample (with each pupil weighted equally).

A truly random sample provides an exactly equal chance for each student to be selected in that sample. This type of sample is relatively easy to obtain within a building (list every student in grade 4 and randomly select 25 percent for example), and perhaps could be managed in a school district with only a few buildings (list every student, regardless of location in building or class, and randomly select a sample). This type of sample is hard to manage in a large school district and is essentially impossible at a state level (to list every student in a given grade in every building in the state and randomly select names).

This situation leads one to a *multi-stage sampling approach*, whereby school buildings may be selected prior to random selection of students. As soon as there is an intervening stage before selecting students, such as buildings, it is essential to move to a probability sample. For example, if, within a given state, one school building houses 200 fourth graders while another building houses only 100 fourth graders, one cannot simply choose one of the two buildings at random to be in one's sample. The first school must have two chances of being selected versus the one chance given to the second school because the first has twice as many students in the population being sampled (fourth graders in the entire state).

Accurate, actual probabilities of inclusion must be assigned

*Each Indian student would "represent" 10 Indian students (a 10 percent sample). Each non-Indian student would "represent" 20 non-Indian students (a 5 percent sample).

to every unit (building, geographic area, etc.) to be sampled in any multi-stage sample. *Any sample that makes some prior selections before getting to the actual selections of students is a multi-stage sample.* Once a person has a list of students in a given class, their sample is drawn randomly (unless a specific sub-population, like American Indians, is being oversampled).

Understanding just a few basic principles of sampling (probability, stratification, weighting) will enable a director of assessment to communicate with sampling statisticians and will help to keep from accepting a casual sample that could be designed by a less than knowledgeable person in this very complex and very important aspect of any assessment program.

Another aspect of sampling that is proving to be of great value to education is *matrix sampling.* As applied to educational assessment, it is a technique that permits one to secure a great deal of information about pupil outcomes without requiring individual pupils to spend an undue amount of time being tested. Instead of asking one student to take three hours of testing, one might ask three students to take one hour of testing each, assigning the items or tests randomly to the students. *In a sense it is like selecting a series of separate, non-overlapping samples of students, with each sample taking a different subset of the total number of items to be administered.* Matrix sampling is an integral part of the National Assessment plan. It would not have been possible to collect all of the information that National Assessment collects without using this technique. A typical National Assessment set of items for one age level in a given year would take about 450-500 minutes of time for one student to answer. It would be impossible to ask schools for that much student time for each student sampled. But it is possible to divide the items into about 12 packages (tests) that take about 45 minutes each (including administration time). That way each of twelve different students can take one package, and the total set of items can be covered with a large enough sample of students to produce an acceptable number of responses for each package. *It is a technique to gather a large amount of data without burdening any one person or any one school with excess testing.* As such, it should be seriously considered in any large-scale educational assessment program.

Principle: *A director of state assessment should familiarize himself with basic princi-*

ples of sampling, but must turn to sampling statisticians for development of a specific sampling plan.

Collecting Data

It would be a serious mistake for a director of assessment to gear an assessment program to traditional, well established data collection techniques without at least considering alternative methods. A conscious decision must be made that the methodology that is adopted is the best that can be afforded and that it will not introduce unnecessary unreliability into the results.

Most assessment programs will end up by relying upon traditional paper-and-pencil, group-administered data collection techniques. Such methodology has proven to be quite satisfactory for most cognitive, skill-oriented assessments. However, if objectives call for psychomotor assessment, or certain skills assessment, or attitudinal assessment that requires observation, it may be necessary to utilize individually administered tests or interviews for some portion of the data collection. Individual administration generally requires special training and is invariably more expensive than group assessment. Yet *some objectives cannot be assessed by paper-and-pencil tests.* A choice must be made between ignoring certain objectives or spending the money necessary to gather usable information. Small sample techniques may be particularly useful for data collection that is quite expensive.

The technique of *using taped directions for group administered tests,* as used by National Assessment and others, *should be seriously considered* for all similar situations. It offers a technique for guaranteeing partial uniformity of presenting instructions that is not possible when using local administrators, teachers, and counselors—who vary from district to district and from building to building.

Another decision to make in data collection is whether to use local, voluntary test administrators at all, or whether to use a smaller "professional" group of test administrators. The term professional implies that a small group of persons has been trained specifically to handle all administrative details of a particular instrument. It is not implied that there are professionally trained test administrators that a person hires for this task.

National Assessment uses a full-time national field staff of

district supervisors who make all local arrangements and do some of the actual administration. They are supplemented by locally-trained test administrators who are hired on an hourly part-time basis. Substitute teachers and others with free time are recruited for this role. Maine used a part of the regular professional staff of their state department of education—who were trained by a contractor and then went from school to school to do the assessment. An advantage of this procedure, in the opinion of the director, was to lend credibility to the entire assessment program. Such a plan as used in Maine is costly in terms of professional staff time, but that does not show in an assessment budget. *A plan that calls for trained test administrators who travel from site to site is most appropriate for small sample assessments.* Every-pupil testing would be prohibitively expensive by this method in most states.

The time or times of year for assessment must be decided. Michigan did its assessment for the first four years in the month of January. This, plus the renorming of a new test each year, meant that results of their every-pupil testing were not returned to the schools until the spring. Michigan now has moved to October testing, in order to have time to get results back to the schools while at least half of the year is still ahead. Michigan uses classroom teachers for administration and, thus, finds it feasible to do it all at once (within a given month, but not on a specified day in every district). National Assessment uses a continuous assessment plan: age 9 in the fall, age 13 in mid-winter, age 17 in the spring, and young adults throughout the year. This plan spreads administration throughout the year for the permanent field staff.

If every-pupil testing is elected, the locus of administration may be either the classrooms, an auditorium, gymnasium, cafeteria, or other large area. The choices are the same as those for deciding where to administer a standardized test. Of course, if a decision is made to limit the size of group administrations (National Assessment's upper limit is twelve), then rooms must be found that will be most appropriate. In some instances, a school may have to send some students home in order to free some space for assessment. Such decisions will have to be made on a local basis—keeping in mind that one should develop the most advantageous testing conditions possible (adequate space, proper heat and lighting, good desk or table space, outlet for recorder, if used, etc.).

Specific details of data collection within a given school

building should be developed with one thought in mind—*testing conditions should promote accurate data collection by providing a situation that contributes to positive motivation for taking the assessment instruments.* The whole purpose of an assessment program hinges upon the collection of reliable, and valid information about student outcomes. Data collection procedures should be developed with that goal in mind. Those persons who develop the procedures should constantly be alert to keeping extraneous errors (errors due to poor data collection) at a bare minimum.

Processing Results

The *first major task of processing is the scoring of the assessment questions.* Those which can be machine scored certainly should be, probably by one of the large high-speed document readers that are designed to handle such tasks most efficiently. If a state department of education has its own document reader, it probably will elect to score assessment tests on it. That may or may not be the most efficient approach, but it may be the only politically viable approach.

It used to be that large scale testing implied the use of separate answer sheets routinely. That is no longer obligatory since *several document readers exist which can score test booklets themselves.* If assessment is to be done for young children, this is an important consideration.

For assessment questions that cannot be machine scored, that require either clerical or professional reading, other scoring procedures will need to be developed. If item development has been carried through thoroughly, there will be a specific scoring rationale and answer(s) for each question. For the short-answer type of question, a well-trained clerk will be able to learn a key, score, and code the answers (for data processing and storage). For longer answers, such as essays, or taped singing, or art works, it will be necessary to secure the services of professional subject matter specialists who can evaluate the adequacy of each response. These *professional scorings can be very time consuming and very expensive.* Yet for some objectives, the only questions that seem to measure them may require professional scoring. National Assessment tape recorded the singing of respondents, for example, and has had them scored by "listeners" who are music teachers. To assess such a skill there was no alternative approach. National

Assessment either had to pay the price of data collection and scoring or decide that it did not want to devote resources to its performance objectives in music.

Professional scoring immediately raises problems of scoring reliability. It is a serious problem, but it can be handled reasonably well with adequate training of scorers. Scoring reliability in such situations will not be at the same level of accuracy as with multiple-choice items, of course. Adequate training usually means an iterative process of: (1) learning the method, (2) scoring, (3) comparing scores with other professionals, (4) discussing and resolving differences, (5) scoring again, and (6) comparing and discussing again; until all scorers are producing comparable scores on the same responses. The process also requires periodic quality control measures so that an individual scorer does not develop any idiosyncratic series of responses as time passes. Specialized consultant help from persons who have done this type of scoring may be needed to get a system underway. This task may be contracted to an agency with experience in doing it, or it may be managed in-house. Probably a director of assessment never would have the regular staff do the professional scoring, since it should be done by professionals in the subject area being covered.

Long before actual scoring gets underway, it will be necessary to develop a system for storage and retrieval of the results. It is not appropriate to approach a data processing office with results in hand. *The time to approach data processors must be at least as soon as items are ready for packaging together into tests,* and preferably while scoring rationales for individual questions are being developed. The point of this suggestion is that it may be possible to avoid or at least minimize some of the problems of real-life data handling by having an expert data processor point out the potential problems long before facing them. An innovative item writer may develop some elaborate process for recording responses that could be fine for his purposes, but could create unnecessary problems for a programmer. Such an "innovation" should be retained only if it gives clear promise of providing information not readily available otherwise. If such an innovation is simply a personal preference of an item writer, and if it is not practical for easy data processing, it should be eliminated at the item development stage, rather than later when twenty other items have been developed in the same format.

Please note that *it is not being suggested here that a data pro-*

cessor should be given veto power over the creations of innovative item writers. What is being suggested is that a director of assessment should weigh very carefully the practical problems associated with each data collection methodology along with the potential outcomes. To do this job, he needs the input of a good computer man at each point that an assessment decision may relate to scoring, data storage, and data analysis.

But a good data processor will be more than a monitor of potential problems. He is very apt to have innovative suggestions himself for possible changes in data collection procedures that will produce even better information than would have been gained otherwise. Each specialist in the total assessment process has the potential of improving any aspect of the program and input should not be confined exclusively to an area of expertise. A director of assessment is well advised to have all instrument development and data collection procedures reviewed carefully by a good computer consultant—who also realizes that not all assessment decisions are his to make.

A major task of data collectors and processors will be to *develop quality check points at all levels of data handling.* From the time that a student makes a mark in a test booklet or on an answer sheet—through the coding that may be done by a scorer—through the on-site checking of answers—through the mailing and receiving of booklets—through the scoring agency's checking procedures—through the running of answer media through a machine—through the detection of scoring abnormalities—and finally into a data storage system, there is the need for constant checking of the sheer accuracy of each process. There is no need to spend $10, or $20, or $30, or more for a good test item if results obtained from it are going to be handled carelessly. *Errors due to data handling can and must be reduced to miniscule levels.* Technology exists for doing this. The major ingredient necessary to the task is a commitment to see that it happens, followed by adequate resources, including personnel.

If data storage and processing are done under contract to an external agency, a director of assessment should have a competent staff member or consultant to monitor the process, particularly if anything other than routine multiple-choice testing is to be done.

In like fashion, there is a need to plan carefully and far ahead of time for the data analysis itself. If anything, *a data analyst should be involved in assessment planning even before a data*

processor. This may seem to be putting the cart before the horse, but it isn't. *A data analyst must know why he is analyzing results in order to decide best how to analyze them.* Thus, data analysis planning must be an integral part of assessment planning. A good way of bringing an analyst into the planning process is by providing him with a set of specific questions that assessment results are supposed to answer. By knowing the questions, he can plan his analysis to provide the appropriate information, or he can point out that the data that is to be collected will not provide the desired information. It is far better to know at the beginning of an assessment project whether a certain analysis can or cannot be done, rather than after data are collected. *An analyst can analyze data; he cannot create it.*

A good data processor or analyst will have a long series of specific judgments to make that relate to the details of a given assessment program. Selecting a specific computer system is most important if the analysis provided is to be cost effective; using the computer of a nearby university may or may not be cost effective; using a computer in a distant city may or may not be cost effective. *To determine cost effectiveness, one needs to consider such variables as rates, computer capacity, access time, and other things that are quite specific.* A small assessment program may be handled quite easily by a small computer set-up, and it might, in fact, be more expensive to use a larger system. A large assessment program or one that is doing very innovative analyses may need to be tied into a much larger, more complex system. The principal point of all this is that a director of assessment and/or his analysts need to weigh alternative computer systems rather than routinely select the one that is closest geographically. National Assessment, for example, is located in Denver and is using computers in Princeton, New Jersey and in Kansas City.

A good analyst will need to *develop a system* for producing the routine reports *that will form the basis for assessment reporting over the years.* He may or may not be able to do this prior to the first reporting of results. Assessment planners themselves may not be able to make final determinations of modes of reporting until after they have received feedback from users of results. Thus, the first efforts of an analyst may be altered over time, but his objective will be to routinize that portion of the analysis that is to become a routine part of the reporting. This task can be

handled by the assessment staff (if they have the skills), by consultants, or by a contractor. Once established, it will need only to be maintained and related to potential new areas for data analysis each year. Most directors of assessment will want to secure good consultant help to establish the system, but will not need to keep an innovative analyst on staff to handle routine ongoing analyses.

A state assessment program will need one or more top notch staff analysts only if it is committed to "real" analysis versus "routine" data runs. Reporting of percentiles routinely for all pupils, reporting of school district means and standard deviations, and reporting of item p-values do not, of themselves, require high level analytic skills after a system is established. But *attempting to search through the data for relationships or breakdowns* that are not a part of routine reporting *calls for a continuing commitment to analysis* in addition to a commitment to reporting.

In a sense, it is being suggested that a state assessment program is very apt to collect data that is amenable to interesting and profitable analysis above and beyond the regular reporting that is planned. Rarely, if ever, is provision made for such additional analysis and reporting; rarely, if ever, is staff time available for it. Ideally, an assessment staff should include an analyst with ample time and a free hand to "milk" assessment data for information not programmed for. But even without one person assigned to this task, there are several other ways to get at it. A director may specifically release a staff member for some period of time (a month or two) to seek additional information from the data. He may encourage professors in universities to do analyses on assessment data and/or to encourage graduate students to do so. He may hire an "analyst in residence" for a year to do nothing but work for additional information.

While the nature of a specific assessment program will determine its potential for analysis, the reason for these latter suggestions is the author's belief that *in education we collect much more data than we ever utilize*, and that *we are missing opportunities to analyze data beyond the routine analyses* that are planned for in any project.

Reporting

For purposes of this discussion, reporting will not include the actual preparation of summary statistics and tabular results from

assessment data—which is considered to be a part of the analysis procedures. Reporting will be considered to cover dissemination activities (printing and distribution of reports), interpretations, work with the media, work with users of assessment results, conferences or workshops on assessment, etc. *Reporting, then, includes all those activities that are concerned with distribution, interpretation, and utilization of assessment results.*

The *major decisions* to be made in any reporting plan are *what to report and how to report.* The simplest decision to make is to say that the summary statistics and tables produced by one's analysts will be printed and sent to a selected list of persons. To some persons this is what reporting is. An assumption underlying this approach is not only that the potential users of assessment results have the capabilities to read, understand, and interpret statistical summaries, but that they have the desire to do it for the assessment results. Certainly there are some users of assessment results who do satisfy these assumptions, but they comprise a very, very small audience, made up primarily of educational researchers, plus a few others.

Most reports from large scale assessments have gone beyond mere production of tables of results, but not very far beyond. The most typical approach seems to be to add a few "extras" to the basic statistical reports, such as:

- Brief explanations of why and how assessment was carried out;
- Brief explanations of how to "read" the tables of statistics;
- News releases that consist of translating some of the numerical results into words; and
- Public reporting sessions that attempt to cover the first three points in a verbal fashion.

For the most part, these efforts seek primarily to help potential users of the information to understand the statistical reports. For the most part, they do not interpret the results, nor attempt to suggest why they are as they are, or what one should do about them.

Many, perhaps most assessors feel that their job is basically complete when they have produced and distributed reports which can be read and understood (hopefully) by potential users of the results. Let's assume that that is an assessor's role; how should he carry it out? The "what" to be reported was decided when decisions were made earlier about reporting categories for the results, the specific summary statistics to produce, and the type of scale or scales to use (item p-values, district percentile norms, school means and standard deviations, etc.). The basic reporting decisions relate to specific formats for the presentation of results—types of tables, use of graphs, etc. In addition, it is necessary to prepare some minimum explanatory materials with some examples of reading the tables. News releases need to be prepared and it may be important to set up one or more press conferences.

This type of reporting does not require large blocks of staff time. If analysis has been contracted to some external agency, undoubtedly they will have been asked to include sufficient explanatory material so that their reports can serve directly as the assessment reports. In such a situation, a director of assessment need only be sure that appropriate press releases are prepared and that the reporting obligations will be taken care of. If reports are prepared in-house, the same analysts who prepared them may be asked to include sufficient explanatory material to make them readable, or the director of assessment or some other staff member may do it. The point is that very little regular ongoing staff time is needed to prepare such reports. Certainly a director could not keep one staff person busy doing such tasks, unless an assessment program is so large that masses of results are being printed on a year-round basis.

If one broadens the definition of reporting to include activities beyond production of readable reports, then the whole process may take on quite a different character. For illustrative purposes we will now move from a minimal reporting plan to a very elaborate one, without discussing the multitude of in-between possibilities. The reader can make up almost any type of in-between plan by adopting some portion or portions of a "total" reporting plan.

> **Principle:** *A "total" reporting/utilization plan consists of any and all activities that are deemed important in order to be assured that*

> *each audience for which the assessment has been designed does, in fact, receive information (both results and interpretations of them) which helps it toward better educational decision making and implementation of positive changes in education.*

A total plan requires a commitment to action on the part of the agency doing assessment. If that agency is a state department of education, it means that the CSSO and his Board are willing and anxious to see that assessment results actually are utilized. It means that considerable resources will be devoted to this task. It assumes that action is sure to take place only if someone is charged with responsibility to see that it does take place. It assumes that *publication and distribution of reports is not a sufficient program to assure that action will take place.*

The beginning of a total reporting/utilization plan, then, must be *a commitment to seeing that something happens in the schools of the state (or other area) as a result of the state assessment.* This means a commitment to *serve as a catalytic agency* to see that interpretations of assessment results do take place by, with, and for the various assessment audiences. This commitment does not mean that state department personnel will be interpreting assessments results and making unilateral decisions that reading instruction is bad, and that it can be improved by doing X and Y and Z.

Let's assume, for illustrative purposes, that a state assessment has been completed in the state of Limbo in mathematics at grade four. And, let's assume further that a state has used criterion-referenced tests that relate specifically to 40 different performance objectives in mathematics. Let's assume that each objective was measured by five items of a paper-and-pencil type, that every fourth grader in the state took a common core of fifty items (five each for the 10 most important objectives) *plus* ten additional items (five each for *two* of the additional 30 mathematics objectives) on a statewide sampling basis.

If there are 30,000 fourth graders in this hypothetical state, then there would be 30,000 results for each item for each of 10 basic objectives, as well as 2000 results for each of the other 30 objectives. Let's further assume that one large city contained 7500 of all the fourth graders in that state. For that city there would be

7500 results for the 10 basic objectives and 500 results for each of the other 30 objectives.

Now what might a total reporting/utilization plan include:

1. A report of statewide summary results for each of the 40 math objectives.
2. A report of statewide summary results for each of the 200 questions.
3. A report of individual school district results for each of the 20 basic math objectives.
4. A report of individual school district results for each of the 50 basic questions.
5. A report of individual class results for each of the 10 basic math objectives.
6. A report of individual class results for each of the 50 basic questions.
7. A report of individual pupil results for each of the 10 basic math objectives.
8. A report of individual pupil results for each of the 50 basic questions.
9. A report of school district results for the one large district of the state for all of the 40 math objectives.
10. A report of school district results for the one large district of the state for all of the 200 questions.
11. Review at the statewide level of all objectives that fall below reasonable expectations. Consideration of whether texts and/or other materials are adequately covering those objectives and of whether fourth grade teachers are adequately prepared to cover those objectives. This could lead to revision of materials or preparation of new materials. This could lead to a series of workshops on instructional methodology for fourth grade math. Implementation through statewide committee of teachers and math educators.

12. Review at each school district of each of the 10 basic objectives that fall below expectations. District-wide involvement in improving materials and/or methodology needed to improve performance on the deficient objectives. Implementation through a district-wide committee of teachers and math educators.

13. Provision of supplementary instructional materials for each of the basic objectives, to be made available to individual teachers, with subsets of questions for each objective for a teacher to use after special instruction. Implementation by each fourth grade teacher.

14. Provision of specific individual student assistance for objectives for which a given student scored below expectations. Implementation by each fourth grade teacher.

15. Initiation of item (11) by state assessment staff and some monitoring of items (12), (13), and (14) via workshops around the state and possibly by reassessment of selected objectives later in the year or the following year.

16. Reporting of the results of (11) to the state board of education and/or legislators and/or the media in non-statistical language. Emphasis on reporting to non-educators would be on what the results mean, not what they are. Interpretations could come from those professionals plus competent lay persons who can read assessment results easily and relate them to curriculum and to instruction.

17. Appointment of a committee to review the total mathematics situation in the state in light of the assessment results—Were objectives realistic? Were items appropriate? Is

mathematics instruction in this state at a satisfactory level? Is teacher training of fourth grade teachers in mathematics at a satisfactory level? Are there implications for public policy? Is there any way to relate these results to neighboring states or to the nation? Such a committee should be a small working committee of math educators, teachers, and lay persons who can call on others for consultant help.

18. Encouragement of the state mathematics teachers' association to attempt the same tasks as in (17), or to do it jointly with the state department.

19. Asking the committee in (17) or a special ad-hoc committee to consider additional analyses of the mathematics results. Do students who used testbook A do better on some objectives than those who used testbook B? Do low SES students have particular problems with certain objectives? By identifying school districts or even teachers where students do particularly well, can one isolate any elements of instruction that may have helped to produce the results? Are certain results more closely related to general scholastic ability or general reasoning than anything else? Is age alone related to any of the results? To what extent does "liking" mathematics relate to the results? Many additional questions could be raised. Ideally, such a committee should have posed these questions prior to data collection so that information could be collected routinely at assessment time. But much of this information would come from principals rather than students, so it could be collected even after assessment.

20. Development of specific hypotheses to be tested that relate to questions investigated

in (19). Assessment results may give cues to relationship, but further study may be necessary if one seeks answers rather than cues.

21. And others.

The reader could go on to think of very specific actions that might be taken to attempt to tease information out of the mathematics assessment results in this example. Different assessment data collection would lead to quite different actions. The major point of this example is that *there are a great variety of actions that can be taken under a total reporting/utilization plan for any assessment.* In fact *they must be taken if it is to be a total plan rather than just a printing and distribution plan.*

A total reporting/utilization plan will require the active cooperation of large numbers of professional educators and lay persons within a state. The best persons to interpret assessment results are persons with knowledge in the given subject area, with teaching experience, and with a desire to do the job. The assessment staff can provide technical expertise to committees or individuals working on interpretation. It should be obvious by now that the author believes that *interpretations must be built into a total plan; one dare not simply assume that it will happen.*

An important principle here is that interpretive reports should be geared to their specific audiences. A report to fourth grade teachers whose students took the math questions will be very detailed; a report to the state board of education on math results statewide will summarize them very briefly and then suggest what needs to be done to improve the situation. *Enough information should be given in each report so that a skeptic may question the interpretations and go back to basic data if he desires.* Lay persons, particularly, need help in relating assessment results to "expectations." Knowing that 92 percent of all fourth graders achieved objective seven makes more sense if one also knows that teachers feel that 90 percent achievement is quite satisfactory.

The written word is only one means of reporting assessment results. The sheer volume of professional literature that crosses the desk of every teacher, every school administrator, and every professor of education augurs against any expectation that assessment reports automatically will be read. To be confident that assessment

results will be attended to by some significant number of educators and others, the audience must either be required to read the reports or listen to them, or the reports must reward their attention by presentations (written and oral) that command interest by themselves. Communication specialists can provide a host of methods to promote attention to and interest in assessment results. A few selected examples are given as follows:

- Slide presentation of statewide results, developed by the statewide committee.
- Specific reports that focus on one contributing variable.
- Public reporting session of statewide and/or district results.
- Reporting sessions at meetings of service clubs and women's organizations.
- State department sponsored reporting sessions that require, as an obligation of attendance, that each attendee organize at least one local reporting session in his own community.
- State department sponsored workshops on state assessment geared to county or intermediate district personnel.
- State department sponsorship of follow-up research studies geared to assessment-developed hypotheses and carried out in local districts.
- Securing time on the conference programs of the various professional conferences held in one's own state. Inviting lay persons to attend.
- Securing time on a TV talk show for a panel of educators to discuss the statewide results.
- Developing a "do-it-yourself" assessment test for adults, in order to show lay persons what their children can and cannot do. Such a test could be used at public meetings of many types, or could be administered on TV.

Dozens of ideas can be generated that could make up an "active" dissemination program. Such a program would, of course, cost quite a bit of money and require a regular ongoing staff of disseminators. National Assessment has a staff of three full-time professionals, plus support staff, working on dissemination alone. They could use more. A large state could use a staff at least that large and even a very small state would require at least one full-time professional person to even begin to do an adequate job of dissemination. A lot of the work is organizational and promotional in nature. A disseminator would have to rely on LEA personnel, university personnel, and lay personnel for much of the actual interpretive work.

An active dissemination program can begin in a small fashion and then expand as resources become available. From a staffing point of view, it has the advantage of being very flexible. No state could possibly attempt everything that could be done; any state can attempt some of the things that could be done. *A person working actively in dissemination must have the quality of being able to get others interested in and working actively to interpret and promote interpretation of assessment results.*

Starting an active reporting/utilization program may mean working actively on a day-to-day basis with one or two or three school districts only, in order to try out ideas and to demonstrate what a local district can do. If such efforts are successful, a disseminator has a good case for securing more help to make the effect statewide. An active reporting/utilization program is limited only by the creativity and the time of the staff assigned to the task.

A side benefit of an active plan will be the regular feedback into the total assessment program of suggestions for improvement. It would be impossible to involve large numbers of people in using assessment results without their requesting additional information or information presented in a different fashion, or suggesting that some things be dropped because of lack of utility. *A good assessment program will be a dynamic, ever-changing program of information gathering:* to keep it responsive to the needs of people for information about education, one needs constant feedback from the consumers of its results.

Principle: *A large scale assessment program needs a well developed, detailed, well-staffed*

program for reporting and utilization of its results. It must not be assumed that good things will happen to education just because good information is being provided to people: it is necessary to make the good things happen.

Chapter 5

EVALUATING ASSESSMENT

Any large-scale assessment program should have its own internal performance objectives, at least informally, which can be evaluated in an effort to improve the program. After the initial year of operation, and periodically thereafter, a director of assessment and his major advisers should review each aspect of their assessment program and decide whether it is still appropriate. Some of the benefits from this procedure are:

- After securing data and seeing it interpreted and used, a director is in a better position to decide if it really is the information that is needed most from assessment.
- After reviewing objectives and questions in light of the information attained, he is in a better position to decide if those objectives and exercises are getting at what is wanted.
- After getting feedback from schools on the entire data collection procedure, he is in a better position to improve.
- After seeing which analyses of results have been most useful to people, he is in a better position to plan for additional and/or different analyses.

There is no real substitute for hindsight. A director of assessment should plan specifically to devote time of staff and of his policy advisers and of himself to looking back over his program in an effort to seek places where changes will improve the program and move it in the direction of providing ever more useful information to educational decision makers.

Checkpoints for a Quality Assessment Program

A quality assessment program can be identified through certain significant characteristics. Some of the characteristics are as follows. The program:

- Has carefully delineated goals that are understandable and important to a known audience.
- Has a realistic number of goals that are attainable with existing methodology.
- Has established priorities among its goals and devotes its major efforts to its major goals.
- Is not neutral: it is designed to gather information about what is considered to be good and/or important in education.
- Is based upon a plan that includes specific objectives which it is striving to attain.
- Utilizes sampling techniques designed to provide results at a level of accuracy that is acceptable to users of results.
- Uses data gathering instruments that *do* measure the objectives of that assessment.
- Collects data in a manner that introduces a minimum of error into the data.
- Scores and processes results in a manner that introduces a minimum of error into the data.
- Uses analytic techniques that provide those types of breakdowns that are needed by decision makers, but also attempts to tease out useful information not necessarily a part of the original design.
- Reports results in ways that are easily understood by its various primary audiences.
- Provides help in the interpretation of results and their implementation for the improvement of education.
- Is staffed by persons who have the skills and abilities to develop and direct a high quality program.

- Provides for the active and widespread involvement of representative groups of persons who can speak for the interests of all major audiences for assessment results.

- Makes provision for securing high quality consultant services, either within or outside that state, for any and all areas that can profit from such services.

Ingredients essential to any quality assessment program are *common sense, flexibility,* and *hard work.* An assessment staff must weigh human concerns against technical criteria for assessment, must be willing to change direction whenever a better option is uncovered, and must be dedicated to the importance of their own efforts.

By the end of the first year of operation of an assessment program, the director and staff probably will know how to design an effective quality assessment. *The best state assessment programs are those that are tailored to the most important felt needs for assessment information within that state.* Thus, no one model can be best for everyone. If each decision that is taken in the process of developing a specific state plan is made on the basis of maximizing the quality of assessment results, then a good quality plan is certain to emerge. If compromises must be made, because of lack of resources, they should be made primarily on quantity of the program (fewer areas, sampling of content, elimination of questions that require expensive data collection methodology, etc.), but not on its quality. One is better advised to have ten bits of useful, reliable, pertinent information than to have 100 bits of questionable, unreliable, irrelevant information. There is no substitute for quality.

Assessment is one component of accountability. It is a vital component, but not the sole one. It provides information that is essential in the process of determining whether the young people of this country are acquiring the knowledge, the skills, and the attitudes that our society deems essential for its continued growth. The basic purpose of any large scale assessment program is to effect positive change in society by broadening our base of information about educational outcomes.

Chapter 6

ASSESSMENT IN OPERATION

In order to illustrate assessment "in action," the programs of five different states are presented here. These states were selected as being diverse on several dimensions: purpose, instrumentation, data collection, etc. Their diversity relates primarily to differing purposes. No one program is necessarily better or worse than the others. Each one contains interesting variations that may be useful to persons who are planning assessment or revising their plans.

Minnesota assessment is designed to gather summary information across the state: *Michigan* assessment is designed to gather information for each pupil in its defined population. *Pennsylvania* assessment is a voluntary service program: *Florida* assessment is required for all LEAs: *Texas* uses a sample of school campuses (buildings) from across the state. Because of their varying characteristics, the assessment programs for these five states are summarized here.

In order to provide uniformity of these summary reports, a 30-item questionnaire was used to secure information. The questionnaire is presented first, and each summary follows the outline used in the questionnaire.

State Assessment Questionnaire

Purpose

1. What is the purpose of your assessment?

2. Are general objectives for your state assessment available?

3. What decisions or types of decisions will be geared to your assessment results? What are the major anticipated applications of the results?

4. What are the major audiences for your state assessment results?

Administrative Structure

5. What is the administrative structure for your assessment?
   (5.1) Is there an advisory and/or policy committee or committees?
   (5.2) Is there a director of an assessment staff? Please outline total staff and where you are located within the SEA structure.
   (5.3) Are there regular or occasional advisors and/or consultants?

Instrumentation

6. What type(s) of information is your program primarily interested in? (Cognitive; affective; psychomotor)

7. Are you using a norm-referenced or criterion-referenced approach or a combination?

8. What educational area or areas are being assessed? Are they the same every year or on a cycling approach? What cycle?

9. Are there specific stated objectives for each educational area? Are they available to others?

10. Who developed the objectives? Explain the process briefly.

11. Who developed the questions used in your assessment? Did you use existing tests or other questions?

12. Are the questions all multiple-choice?

13. Were questions tried out before use in your assessment?

14. How were questions selected for actual use in your assessment?

15. How are questions packaged into tests? Are they like conventional published tests?

Data Collection

16. Do you assess a total population or do you sample?

17. Please define your assessment population.

18. If you sample, please outline procedures. Do you stratify? Do you select districts or schools or pupils or all three?

19. Do you do item sampling as well as pupil sampling?

20. Explain data collection, e.g., group or individual; performance or paper-and-pencil; who administers instruments?

Data Handling

21. How is scoring done?

22. How is data processed and stored?

23. What are your reporting categories? (age, sex, region, size of school, etc.)

24. What types of scores do you use for reporting?

Reporting Results

25. Please explain your reporting process.
    (25.1) Are printed reports of results available?
    (25.2) For whom?
    (25.3) Are interpretations of results made? If so, how?
    (25.4) Do you use the media to publicize results? If so, how?
    (25.5) Do you hold public meetings to report and/or discuss results?

26. Do you carry on any utilization and/or application efforts that go beyond reporting?

Program

27. What is a typical 12-month schedule for major assessment activities?

28. What is the total cost of your assessment for one year? Does this cover all staff personnel who work on assessment?

29. Do you evaluate your own state assessment program?

30. Please explain any important points about your assessment that have not been covered above.

Minnesota Statewide Educational Assessment Program*

Purpose

1. The purposes or goals of the Minnesota Statewide Educational Assessment Program (MSEAP) are:
   (1.1) To provide the educational decision-makers in Minnesota with a program to diagnose relevant strengths and weaknesses in Minnesota schools and, subsequently,
   (1.2) To provide resources at the various levels of indicated need for purposes of ameliorating weaknesses and capitalizing upon the strengths within the educational systems of the state.

2. The specific objectives assigned to the program are as follows:
   (2.1) To determine the level of performance of students in this state in the cognitive, affective and psychomotor domains.
   (2.2) To identify the variables which account for the variations in student performance.
   (2.3) To report the results of this investigation to educational decision makers in the Executive and Legislative branches of state government, the State Board of Education, the Department of Education, local school administrators, local school boards and interested citizens of the state providing a guide for the allocation of school resources.
   (2.4) To longitudinally report the extent to which progress is being made in Minnesota schools toward improving student performance within the State of Minnesota.

*Prepared by John W. Adams, Director, State Educational Assessment, Department of Education.

These purposes and objectives were presented to, and approved by, the State Board of Education on October 4, 1971. No legislation has been mandated by the Legislature or requested by the department. However, the department is requesting financial support for the program from the Legislature through its program budget.

3. Major decisions would relate to the department's management of the educational program from the state level. Additional resources will be secured to work towards the improvement in performance if the results of certain groups of students or schools or school districts warrant such extra attention. Specific remediation strategies will be based upon assessment results and related follow-up research studies.

4. While there are a number of potential audiences for the assessment program, primary targets include decision makers at the state level including program managers within the department, the State Legislature, and the citizenry of the state.

Administrative Structure

5. (5.1) The Assessment Advisory Council is a 25-member body with representation from professional educational organizations, other state agencies, higher education, nonpublic education, the Governor's Office, the State Legislature, citizen groups concerned with education, and representatives from the Department of Education. With the participation of these groups, the various audiences with which the department must communicate are represented to express opinions and identify concerns that surface in these audiences.

Additionally a five member Technical Advisory Committee has been formed to provide technical expertise in areas related to educational assessment, measurement, and evaluation. The membership includes public school personnel with research and evaluation responsibilities and professors from higher education

institutions with particular expertise in research and measurement.

The responsibilities of both advisory groups are limited to general policy establishment; specific responsibilities relative to implementation remain with the Commissioner of Education and his designees within the Department of Education. Each advisory group meets about once every six weeks.

(5.2) The Director of State Educational Assessment reports to the Assistant Commissioner for the Division of Planning and Development. There are five Assistant Commissioners in the Department of Education, each of whom reports to the Commissioner of Education. At the present time the Director is the only staff member in the Assessment Section.

(5.3) Outside consultant help has been secured to assist in the planning and implementation of the assessment program.

Instrumentation

6. The assessment program will include cognitive, affective, and psychomotor performance appraisal.

7. A criterion-referenced approach is being followed utilizing National Assessment released exercises and exercises developed in Minnesota. Other sources of exercises will also be considered.

8. Reading and mathematics are the initial areas developed for assessment purposes, to be followed by literature, social studies, science, art, music, writing, career and occupational development, citizenship, and health and physical education as resources permit. A five-year cycling approach will be used covering at least two areas each year.

9. Specific performance objectives have been developed in reading and mathematics with the other areas to be developed in the future. Materials will be shared with others.

10. Objectives were developed following procedures developed by National Assessment. Three principle audiences were involved:

    - Subject-matter experts (professors, specialists);
    - School personnel (teachers, administrators); and
    - Citizens of the state concerned with education. Extensive involvement from professional groups (math teachers, reading councils) has been encouraged.

11. Professors from the University of Minnesota and specialists within the Minneapolis Public School system were involved in writing exercises for the assessment program.

12. Exercises are primarily multiple-choice, however a determined attempt is being made to go beyond the multiple-choice format.

13. An attempt has been made to secure extensive try-out data before exercises are fully implemented on a statewide basis.

14. Content validity is insured by various review panels before actual use.

15. Exercises are assembled into packages similar to National Assessment procedures. Students mark on the assessment packages. Packages are quite different from standardized tests in appearance.

Data Collection

16. A sample of age and grade members is selected.

17. Minnesota assessment population includes 9, 13, and 17 year-olds as well as students in grades 4, 8, and 11 in public and nonpublic schools.

18. In general, the sample design for Minnesota meets the following requirements:

(18.1) The sample is a probability sample, i.e., each student in a given age, class or grade in a public or nonpublic school in the State of Minnesota has a known positive chance of inclusion in the sample.

(18.2) Each of the ten geographical reporting regions of the state (two of the eleven regions are combined to form a single region) are represented in the sample so that results can be reported for each of them with nearly equal statistical precision.

(18.3) Each reporting variable, or any combination of two reporting variables, can be reported for up to eight reporting groups. This requirement allows for the analysis of interaction effects of certain pairs of reporting variables, i.e., differences in the effects of one reporting variable for different levels of the remaining variable will be examined.

(18.4) If, for a given grade and/or age level, the amount of time it takes to administer the entire set of assessment exercises to a given student is longer than desirable, a matrix sampling approach will be developed to shorten the length of time each pupil will be tested.

(18.5) The school and pupil-school sample sizes will be such that estimates of the sampling variability of the reported results can be estimated from the sample data.

This sample design involves the use of the state's ten reporting regions as a stratification variable to give the sample "representative credibility" by geographically spreading it across the state. Selected school level variables which relate to educational performance are also used as stratification variables within each of the ten reporting regions. Stratifying by these achievement-related school variables (per pupil expenditure and type and size of community in which the school is located) provides for more homogeneous groupings of schools, thus enhancing the statistical precision of assessment results.

The sample design involves two sampling stages. The first stage consists of selecting a random sample of schools within strata; the second stage is the selection of a random sample of students within the schools selected at stage one.

19. See item 4, page 95.

20. Data collection is by group-administered (20 students in smaller schools, up to 40 students in larger schools), paper-and-pencil responses using a paced-tape for control purposes. A trained Exercise Administrator conducts the assessment session. The Exercise Administrator is employed by the assessment contractor.

Data Handling

21. Scoring is done by a contractor, who also prints the assessment packages.

22. Results are processed by the contractor who scores the packages.

23. Reporting categories include age, grade level, sex, region of state (10 geographic regions), size and type of community, school variables, etc.

24. Performance on individual exercises is reported for each reporting category.

Reporting Results

25. (25.1) Printed reports are available in both complete and summary form.

(25.2) They are widely distributed to schools, and are available to anyone including the press.

(25.3) Interpretations are not made by the department. However, we attempt to identify and work with those interested and qualified to make interpretations.

(25.4) The media, particularly the press, has been most helpful in disseminating results.

(25.5) Regional meetings are held around the state to discuss results and solicit interpretations.

26. Additional staffing will allow greater activity in the utilization/application function.

27. Data is collected according to the following schedule:

Fall – 13-year-olds and 8th graders
Winter – 9-year-olds and 4th graders
Spring – 17-year-olds and 11th graders

Reports are completed within 6 months following the collection of the data.

28. Costs for the 1972-73 school year will be approximately $180,000. No realistic per pupil cost can be given at this time because some costs must be amortized over future assessment efforts.

29. No formal outside evaluation of the assessment program has been done or is contemplated. Constant, critical review is provided by the Assessment Advisory Council and the Technical Advisory Committee as well as departmental staff. It would be difficult to conceive of a more rigorous evaluation.

30. Future activities for the assessment program will be largely determined by funds appropriated by the State Legislature. Present activities are made possible, to a large extent, by a grant from the Hill Family Foundation of St. Paul, Minnesota.

Texas Education Agency Assessment Program*

Purpose

1. The legal basis for the Texas needs assessment is found in

*Prepared by staff members of the Texas Education Agency.

Bulletin 693, Texas Education Code, Section 11.26 (a) which states:

> *The State Board of Education shall review periodically the educational needs of the state, adopt or promote plans for meeting these needs, and evaluate the achievements of the educational program.*

2. The general objectives of the Texas Statewide Design for Educational Needs Assessment are:

   - The State Board of Education and the Texas Legislature will have timely and reliable information on educational needs of Texas pupils.
   - The Texas Education Agency (TEA) will have a systematic procedure for identifying priority areas of concern and for assessing the status of Texas pupils in the identified areas.
   - Educational needs of Texas pupils will be alleviated by actions of local school districts with the assistance of the TEA.

   The priority areas of concern which guide the Texas needs assessment come from the Goals for Public School in Texas adopted by the State Board of Education.

3. Several decisions have been based upon the overall experiences gained from the most recent Texas Assessment of Reading and Mathematics. At the state level:

   - the reading and mathematics objectives used for the assessment were reviewed by panels of teachers and other educators.
   - Criterion-referenced reading and mathematics test instruments, similar to those used in the statewide assessment, were made available to schools as part of a statewide guidance program. This decision was based on evaluation results from the assessment that indicated a very postive attitude on the part of teachers towards the instructional

value of the criterion-referenced test reports.

- A study has been instituted to determine the value of pretest information from the use of criterion-referenced testing for evaluation of school programs, especially Title I and Title I Migrant programs.

- A study devoted to renewal of the system used for school accreditation in Texas has used some of the results from the reading and mathematics assessment as one output variable.

- The state mathematics curriculum has been revised to include statements of learner outcomes that are to be achieved in Texas.

- The needs assessment results are being used as the information base for policy decisions on management of the elementary and secondary system of education in Texas.

The Priorities Committee of the State Board of Education is currently studying alternatives for establishing long-term objectives (5-7 years) based on the reading and mathematics results for accomplishment in the public schools. It is anticipated that the results from the 1971 Texas Assessment of Reading and Mathematics will become benchmark information for another cycle of reading and mathematics assessment in the future. Also, the results and the experience in designing and implementing this type of assessment will be used as Texas moves closer to educational accountability.

4. The major audiences for assessment reports have been state, regional, and local educators. Plans are being developed to report to general audiences including legislators, parents, and other citizens.

Administrative Structure

5. (5.1) The assessment plans are reviewed by advisory committees both within and outside the TEA. The outside agency advisory committees are made up

of school personnel and regional education service personnel. The inside agency groups are the Executive Planning Committee, made up of division directors, and Commissioners Coordinating Council, made up of Assistant and Associate Commissioners, the Deputy Commissioner, and the Commissioner of Education. Each assessment is managed by a steering committee representative of major divisions within the Agency which has the responsibility for making periodic reports to the review groups.

(5.2) The needs assessment staff is located in the Office of Planning which is a part of the Commissioner's Office and works with each of the following sections—special education, administration, teacher education and instructional services, occupational education, and technology. The Needs Assessment Section consists of four professionals from varying backgrounds—statistical, curriculum, school administration, and sociology. This section is in the Division of Program Planning and Needs Assessment, and the director reports to the Associate Commissioner for Planning.

(5.3) Ad-hoc advisers are used for:

- recommendations about objective and instrument development;
- data analysis procedures; and
- interpretation of assessment results.

Instrumentation

6. The assessment program in the past has dealt primarily with cognitive measures. Future assessment will include more affective areas, particularly career education, which is presently in the pilot study stage.

7. The first major assessment study managed by the TEA—The *Texas Achievement Appraisal Study*—used norm-referenced instruments. This study's purpose was to assess the academic achievement of Texas high school seniors. The next assess-

ment—The *Texas Assessment of Reading and Mathematics*—used criterion-referenced instruments. This project was used to assess the status of sixth graders in these two subject matter areas. Presently, learner outcomes for career education are being rated by various Texas audiences, i.e., students, educators, parents, businessmen, labor groups, and others. Decisions will be made in the near future about whether to use available instruments or develop criterion-referenced assessment instruments for measuring status of students in career education.

8. The educational areas for assessment are identified from priority areas of concern. These priority areas are named through a review process that includes recommendations from advisory groups within and outside the Agency. The *Texas Assessment of Reading and Mathematics* and the present study for an assessment of career education are products of this system. Long-range plans include assessments based on priority areas of concern that are named, and also, cycling in some areas such as reading and mathematics. As assessment measures are developed in different areas, they will be included in the cycle either as a separate area or in combination with other areas, i.e., combining affective domain measures with other measures for a career education assessment.

9. The *Texas Assessment of Reading and Mathematics* used specifically stated objectives as measurement criteria. The objectives were developed jointly by CTB/McGraw-Hill and the TEA. The measurement items were developed by CTB/McGraw-Hill. The learner objectives used for each assessment are included in the reports and are available from the Agency.

The developmental approach being used for the Assessment of Career Education is different from the reading and mathematics assessment. At the present a series of learner outcomes matched with learner characteristics have been developed by a project called *Partners in Career Education*. The Agency is working in cooperation with Partners to get the characteristics and outcomes reviewed and rated by a segment of Texas

citizens made up of different population groups. Also, a pilot study based on two instruments, the *Career Maturity Inventory* (CTB/McGraw-Hill) and the *Assessment of Career Development* (American College Testing Program) is being conducted. The information from these instruments will be compared to the validated learner outcomes and characteristics. If the measurement instruments cover most of the learner outcomes and characteristics, a statewide assessment of career education will be conducted next year. If these instruments do not measure the outcomes and characteristics to the desired degree, the next year will be devoted to instrument development.

10. The reading and mathematics objectives were developed by agency and mathematics consultants and CTB/McGraw-Hill personnel. Teachers, supervisors, and university specialists reviewed the objectives. Career education objectives will be the validated learner outcomes and characteristics.

11. Questions for the *Texas Achievement Appraisal Study* were developed by the American College Testing Program. For the reading and mathematics assessment, the measurement questions were developed by CTB/McGraw-Hill. The additional questions asked of students in order to make distinctions between population groups were developed by the TEA. The mathematics assessment was based on an existing instrument, the *Prescriptive Mathematics Inventory*.

12. Questions except for mathematics are multiple choice. Mathematics questions were open response. This was possible in this area, because numbers were recorded directly on the answer sheet.

13. Questions for *Texas Achievement Appraisal Study* were the same as those used in a 1967 statewide study. A pilot study was conducted to try out the reading and mathematics questions. This study was conducted in a region of Texas that had a population somewhat similar to the state as a whole. The present study of career education instrumentation has been explained previously.

14. Questions selected for use were based on judgments of TEA personnel, test company personnel, and suggestions from educators from school districts involved in the pilot study.

15. Questions are packaged as conventional published tests. Separate answer sheets are used in most instances. The mathematics questions were answered in the test booklet and then coded.

Data Collection

16. A sample based on assessing a grade level on school campuses is selected.

17. The assessment population has been based on one grade level because of the large number used in the assessment. The selection of sample by school campus causes the larger number. *Texas Achievement Appraisal Study* involved 70,000 twelfth graders. *Texas Assessment of Reading and Mathematics* involved 23,000 sixth graders.

18. The random sampling procedures are stratified on the basis of size of community in which campus is located, geographic location, etc.

19. No item sampling has been done to insure useful information to participating teachers and schools.

20. The data collection is group administered, paper-and-pencil test. Pre-test workshops are held for schools participating in the assessment, and administration is by school personnel—counselors and teachers.

Reporting Results

21. Machine scoring is done by contractor, who prints the tests and manuals and distributes them.

22. School reports are printed by contractor who does the scoring, also preliminary regional and state summaries are produced by contractor. Further detailed analysis is done by Agency

personnel. A part of the agreement with the contractor is the purchase of the scoring tape. Additional analyses are completed by the Agency and reports written.

23. The reporting categories have varied some in each assessment. Primarily the reporting categories are sex, ethnicity, size of community. A fairly detailed index for establishing socio-economic status was used in the *Texas Achievement Appraisal Study*. Because establishing this index was thought to be too tedious for sixth graders, a surrogate index called the "education emphasis in home" was used for the reading and mathematics assessment. This index was based on responses to questions about the amount of television watched, the number of books in the home, and whether or not the family subscribed to a daily newspaper or magazines. The response categories were given weights and the students were divided by high or low educational emphasis responses.

24. The reports on the *Texas Achievement Appraisal Study* were by standard scores. The *Texas Assessment of Reading and Mathematics* was reported by percentage of students that mastered the objectives. The reporting on career education assessment is still tentative.

25. (25.1) Printed reports of results for each school and region are provided individually to participants and state summary results are available from the Agency.

(25.2) Class reports and school summary reports are routinely made available to participating schools for use in instructional planning. School district reports were available on the *Texas Achievement Appraisal Study* because all of the high schools in a selected district participated. More recent assessments based on the school campus sampling process have precluded school district summaries, since schools are not selected as being representative of the district.

(25.3) Statewide reports include basic interpretations produced by the TEA. Regions and schools interpret their own results with the assistance of the contractor

or the Agency, if requested.

(25.4) Various media are used to publicize statewide results. News releases are prepared and all reports are made available to them.

(25.5) A tape-slide presentation was prepared from the results of the *Texas Assessment of Reading and Mathematics*. This presentation was made available to regional education service centers for use. Reports about usage indicates that it has been used with various audiences, but predominantly with educators.

26. Utilization and application efforts so far have been directed toward schools for use of the results in planning instructional programs and to other divisions of the Agency for review of objectives, reviewing and redirecting instructional programs on the basis of assessment results, etc.

Program

27. Data collection usually takes place in October, report to schools and pupils in November. A formal statewide report is completed in the spring. In the present career education pilot study, data collection was in March, reports to school and pupils in April, and decision to be made about the statewide assessment in early summer.

28. The major cost of an assessment program depends upon the number of pupils involved, the types of instruments selected, and the kinds of reports returned to school. The costs of maintaining the assessment section are constant and are not a part of the per pupil cost reported. The contractual cost for *Texas Achievement Appraisal Study* was $70,900.00 or approximately $1.02 per student. The contractual cost for the *Texas Assessment of Reading and Mathematics* was $97,000.00 or $4.22 per pupil. The much higher per pupil cost is because of the type of reports sent to schools and pupils and procedures that had to be used for combining the reading results, mathematics results, and the additional information on each participating pupil. These reports do not include the costs for the additional data analyses, report inter-

pretation, and publication of reports.

29. There is an annual review of the assessment program, but each assessment study has been evaluated separately. The primary focus of the evaluation is the usefulness of the assessment information to schools and pupils for planning their instructional programs. For example, the evaluation of the reading and mathematics assessment has confirmed the usefulness of criterion-referenced test results for classroom instruction.

30. One major purpose contained in the philosophy of assessment used so far is that results must be useful for the school participating in a study. This is why the sample selection is by grades on a campus instead of matrix sampling of individuals. The large number of students, the spread of geographic location, and the almost 100% participation of selected schools have made the samples very representative of the state according to standards that have been developed for the assessment.

    The experience with criterion-referenced testing has been positive and it is probable that future assessment will be based on this kind of measurement if at all possible. The needs assessment experiences are being used as a basis for moving toward more accountability in education for Texas schools.

Pennsylvania's Educational Assessment (EQA) Program*

Purpose

1. Legislation in 1963 required the State Board of Education:

    to develop or cause to be developed an evaluation procedure designed to measure objectively the adequacy and efficiency of the educational programs offered by the public schools of the Commonwealth....The evaluation procedure shall be so constructed and developed as to provide each school district with relevant comparative data to enable

*Prepared by staff members of the Pennsylvania Department of Education.

directors and administrators to more readily appraise the educational performance and to effectuate without delay the strengthening of the district's educational program....

2. See items 1 and 8 for formal objectives. Informally the purpose is to give schools an objective information base on which to assess strengths and weaknesses of the school program. The school—not the district, classroom or student—is the unit of analysis.

3. (3.1) Schools may use Educational Quality Assessment results to pinpoint "needs" as basis for ESEA Title II proposals.

(3.2) Revisions of existing programs in mathematics have been undertaken.

(3.3) Results have been used to "humanize" the faculty, to create an awareness of one's impact on students' attitudes as well as their areas of specialization.

4. School administrators (central office and building level) are primary audience. Control of data after that is in hands of superintendents—to disseminate to faculty, school board, press, and community groups.

Administrative Structure

5. (5.1) Presently there are no advisory committees; the state and national advisory committees have been disbanded.

(5.2) The Division of Educational Quality Assessment has a director plus seven professional staff and three supportive staff. The Division is part of the Bureau of Planning and Evaluation which also encompasses Field Surveys (on-site evaluations) and Planning (including ESEA Title III).

(5.3) No advisers currently.

Instrumentation

6. Primarily affective but cognitive as well (see item 8).

7. Norm-referenced.

8. The *Ten Goals of Quality Education* adopted by the State Board of Education are:

   - Quality education should help every child acquire the greatest possible understanding of himself and an appreciation of his worthiness as a member of society.

   - Quality education should help every child acquire understanding and appreciation of persons belonging to social, cultural and ethnic groups different from his own.

   - Quality education should help every child acquire to the fullest extent possible for him, mastery of the basic skills in the use of words and numbers.

   - Quality education should help every child acquire a positive attitude toward the learning process.

   - Quality education should help every child acquire the habits and attitudes associated with responsible citizenship.

   - Quality education should help every child acquire good health habits and an understanding of the conditions necessary for the maintaining of physical and emotional well-being.

   - Quality education should give every child opportunity and encouragement to be creative in one or more fields of endeavor.

   - Quality education should help every child understand the opportunities open to him for preparing himself for a productive life and should enable him to take full advantage of these opportunities.

   - Quality education should help every child understand and

appreciate as much as he can of human achievement in the natural sciences, the social sciences, the humanities and the arts.

- Quality education should help every child to prepare for a world of rapid change and unforeseeable demands in which continuing education throughout his adult life should be a normal expectation.

9. The objectives are defined by the rationale stated for each questionnaire, the dimensions measured, and sample items which are stated in the interpretation manual.

10. A Committee on Quality Education was appointed from the State Board of Education. This group met with national consultants and held public hearings throughout the Commonwealth. Their recommendations were presented to a committee of 150 civic and professional leaders who approved the *Ten Goal. of Quality Education* which were subsequently adopted by the State Board of Education.

11. Most questionnaires in use were developed by department staff. In some goal areas inventories developed and used by university researchers were adopted. Educational Testing Service constructed the basic skills test; their feasibility study was the source of other items.

12. All instruments are of paper-and-pencil, multiple-choice type.

13. A tryout, or developmental phase, was conducted in 100 schools in 1968 to refine tests and procedures.

14. Questions were selected to measure the dimensions that it was agreed were implicit in each goal. Such construct validity as well as content validity, reading level, and reliability (internal consistency) were the primary criteria for selection of items.

15. Questions are packaged into sections by goal area like conventional published tests. Scorable test booklets were initially

used but separate, machine-scorable answer sheets are now employed.

16. The entire student population in a given grade level is used as a measure of that school. Assessment has taken place in elementary schools (where all 4th grade students are tested), middle schools (7th grade), junior high schools (9th grade), and high schools (11th grade). Effective with the 1973-74 school year, 8th grade students will be tested for both middle schools and junior high schools.

17. The assessment population is all 4th, 8th, and 11th grade students in the participating schools.

18. Stratification is used only to select a representative sample of schools to serve as the normative base. The two criteria used for stratification are wealth (assessed valuation per pupil) and size (number of students in the grade level involved). Because of insufficient funding only a limited number of schools can be assessed in a given year. The district superintendent volunteers to participate and school districts are accepted on a first-come, first-served basis. Volunteers who participated in the past are placed at the bottom of the list and have not been accommodated in their requests for reassessment.

19. No item sampling—each student completes the entire battery.

20. Data collection is by group-administered, paper-and-pencil tests. Administration is by local school district personnel (representatives from each district attend a regional training workshop).

Data Handling

21. Scoring is done by a contractor who also prints the booklets and answer sheets.

22. Data are processed by contractor who supplies raw and weighted data tapes, data cards, and printouts which are stored at assessment headquarters in the State Department

of Education building in Harrisburg.

23. Reports are by school building. No students are identified by name. Included in the report are comparisons of:

(23.1) The school scores to other schools in state (percentile rank in state).

(23.2) The school scores to an "expected school score" computed from a multiple regression analysis of some 40 surrounding conditions. Schools first are classified as high, average, or low on an SES index. Separate regression analyses are used for each category, using whichever variables add at least one percent to the variance accounted for.

(23.3) The school score on the 40-some surrounding conditions to other schools in the state.

(23.4) The percent of students scoring in each of five scoring categories (quintile distributions).

(23.5) The percentage replies of students to selected items.

24. Raw scores for the school are reported with their corresponding percentile rank, other scores as necessary for items 23.2 through 23.5.

Reporting Results

25. (25.2) Three copies of a 15-page computerized report are prepared for each school.

(25.2) The reports are hand carried to the school district superintendent and his designated representatives.

(25.3) A two-man team visits the school district for a half-day to interpret the results to the superintendent and whomever else he wishes to invite—typically the administrators of the district, but occasionally counselors, school board members, and members of the public.

(25.4) Further dissemination is at discretion of superintendent.

(25.5) General findings are published, but districts are not identified by name.

26. Annual follow-up inquiries as to:

(26.1) Why the superintendent volunteered for assessment;

(26.2) Dissemination of the results;

(26.3) Use and relevance of the results; and

(26.4) Suggested changes.

In addition, a workshop was held to demonstrate utilization and dissemination procedures used by districts in the past.

Program

27. Assessment has evolved with changes every year so no one year is necessarily typical. In the 1972-73 school year:

June-August:	Writing and printing interpretation manuals. Preparation and mailing of testing materials. Test development (an on-going activity). Writing computer programs for data analysis and school reporting model.
September:	Regional workshops held to train school district representatives for 5th and 11th grade testing.
October:	Testing of 5th and 11th grade students.
November–January:	Processing and scoring of student booklets. (grades 5 and 11). Test development

February:	Regional workshops to train school district representatives for 7th and 9th grade testing.
March:	Reporting and interpretation visits to 5th and 11th grade schools. Testing of 7th and 9th grade students.
April–May:	Processing and scoring of student booklets (grades 7 and 9).

28. Total assessment costs for 1972-73, including assessment staff and contracting work was $315,000. The per pupil cost (105,000 pupils) was $3.00.

29. No formal, systematic evaluation has been undertaken. However, written and oral feedback from school people, desires of the Commissioner of Education and the State Board of Education, together with much agonizing self-appraisal by the assessment staff have all contributed to yearly evolutionary changes in the assessment program.

30. Future assessment will be only in the spring of the year for grades 5, 8, and 11. The voluntary nature of participation is subject to change which might affect public release of results.

    Investigation and experimentation with criterion-referenced tests and reporting procedures have been undertaken.

Florida State Assessment Program*

Purpose

1. To supply the state's educational managers with information needed to guide their decisions. To evaluate programs through the assessment of pupil performance and behavior.

2. To establish and evaluate assessment procedures.

*Prepared by Phyllis Hawthorne, Research Analyst, State Educational Accountability Repository (SEAR), Wisconsin Department of Public Instruction and James Impara, Administrator, Educational Accountability, Florida Department of Education.

3. The objectives for assessment are stated in #1. We also have the hopes of continuously improving the assessment program. The first year's experience has been used:

   - To modify the assessment program and procedures and provide decision makers with some preliminary information on how students' reading performance compares with the statewide educational objectives;
   - To pinpoint weak spots and curriculum areas in need of modification in order to redistribute educational resources where needed;
   - To provide data for use in the school accreditation program; and
   - To provide an analysis of costs of educational programs to become operative at a later date.

4. State and district level educators; public.

Administrative Structure

5. (5.1) No special policy or advisory committees. A Task Force was formed in the Department of Education following 1970 Legislation which prepared a plan for implementing the Act. (The 1970 Act called for the development of assessment procedures. The 1971 Act mandated an accountability program and implementation of the procedures developed as a result of the 1970 Act.)

   (5.2) Evaluation Section of Bureau of Research (Department of Education).

      (5.2.1) Program Director I---------Secretary III
      - Educational Consultant II
      - Statistician III---------Statistical Aide I
      - Information Specialist II (one-half time)

(5.2.2) District and school personnel
- District coordinator for each of the 67 school districts.
- School coordinator and test administrators in each school where pupils were tested.

(5.3) Outside reading consultant was utilized initially until Department of Education employed one at a later date. Other subject matter consultants within and outside the Department also provide input.

Instrumentation

6. Cognitive.

7. Criterion-referenced tests for 1971-72, but norm-referenced tests may be integrated into the program later. Legislation suggests use of both.

8. 1971-72—reading.
1972-73—reading, writing, and mathematics. Subsequent years through 1977-78—subject areas relating to goal areas of communication and learning skills, citizenship education, occupational interests, mental and physical health, home and family relationships, aesthetic and cultural appreciations, and human relations.

9. Reading objectives for age 7, grade 2, and for age 9, grade 4 are now published in Florida's 1971-72 assessment reports. Reading, writing, and mathematics objectives for ages 8, 11, and 14 for the 1972-73 assessment are copyrighted.

10. 1971-72—Contracted with IOX at University of California in Los Angeles for a catalogue of reading objectives for grades 2, 4, 7, and 10. One Hundred twelve reading specialists and 236 classroom teachers throughout Florida selected objectives to be measured from the catalogue, establishing the grades at which each objective should be mastered, and the objectives which were of highest priority.

Outside reading consultant reduced list to a more manageable size.

Contracted with Florida State University to provide reading objectives for an expanded number of reading-related skills, grades K-12, to be assessed in subsequent years.

Contracted with state universities and school districts to develop objectives in other areas.

With the exception of the outside consultant, the basic procedure is the same. More teachers and lay public will be involved in selecting objectives from catalogues.

11. IOX developed questions to measure each of the statewide objectives for the 1971-72 assessment in reading under the supervision of the Research and Development Section of the Department of Education. Development of assessment items in mathematics and writing and redevelopment in reading for the 1972-73 assessment began in March, 1972. To guide test developers, the Evaluation Section prepared a set of specifications for assessment items as a supplement to the general guidelines supplied by Research and Development Section.

    Questions were developed in first draft form by the catalogue coniractors. Final items prepared by assessment contractors (in 1972-73, Harcourt, Brace Jovanovich).

12. No.

13. Some pre-testing was done for the purpose of determining the procedures to be used in the program; included student responses on sample items to estimate total administration time needed for a larger test, and to determine standard errors that may be expected. Same for 1971-72 and 1972-73.

    1971-72—After the objectives were selected, the test items to measure them were contracted for from IOX. They were evaluated by a review group for content validity on a form prepared by the Evaluation Section and the state reading consultant. The purpose was to determine what items should not be reported in the results because of lack of content validity, serious technical difficulties, not measuring the objective, etc.

1972-73—Items were prepared to measure objectives which were selected.

15. 1971-72—Four different test forms were used for assessment in each of the two grades. Each form measured approximately 1/4 of the selected statewide objectives. Each form was given to a different sample of students. All items measuring a single objective were on the same test form. The second grade test forms had 28, 29, 29, and 30 items respectively; the fourth grade forms had 73, 77, 72, and 69 items respectively. The number of items per form varied so that the forms would take about the same amount of administration time, and some item required more time than others.

    1972-73—Three test forms—many more items due to increase in number of subject areas.

Data Collection

16. Multi-matrix sample.

17. 1971-72—Each school with 12 or more students in grades 2 and/or 4 was included. About 20 percent of the students in each grade was tested, totaling almost 53,000. Included were the Educable Mentally Retarded; children whose predominant language was not English; children who were repeating a grade; and, for the fourth grade test, those who had skipped grade three. Only the Trainable Mentally Retarded were excluded.

    1972-73—Same as 1971-72, but smaller sample sizes: 20,000 in grade 3; 20,000 in grade 6, and 15,000 in grade 9.

18. 1971-72—Multiple-matrix sampling procedure. Do not stratify. All schools with 12 or more students included. The principal, school coordinator, or other designated school official chose the students one week prior to assessment according to a detailed set of instructions.

    1972-73—Minimum school sizes varied slightly, otherwise same.

19. Yes, by test forms.

20. 1971-72—Packaging and mailing accomplished by Evaluation Section.

    1972-73—Packaging and mailing accomplished by assessment contractor, Harcourt, Brace Jovanovich.

    District coordinator selected for each of the 67 districts responsible for handling the dissemination of mailed materials and information to school coordinators and test administrators, and for returning them to the Evaluation Section.

    One school coordinator in each school trained test administrators, handled all preparations at the school, supervised the actual testing, and prepared materials for return to district coordinator.

    1971-72—Test administrators read test instructions and questions to students; collected test books after completion, and filled in pupil data requested on the cover of test.

    1972-73—Same but less reading of instructions; then only grade 3.

    Individual pupils mark answers to multiple-choice questions or free response items.

Data Handling

21. 1971-72—Hand-scored by Software Program & Associates, Titusville, Florida, using answer keys provided by IOX. No partial scores allowed; each item scored as either right or wrong. Contract called for 100% accuracy in scoring. Department of Education checked accuracy. Overall error rate was less than 1%, but items with excessive scoring errors (above 9%) were not reported in the results.

    1972-73—Scored by Harcourt, Brace Jovanovich and National Computer Systems under contract. 99.99% accuracy.

22. IBM 360/40 magnetic tape storage. Storage by school, district, individual student data are kept, but not by student ID.

23. Results are reported only by district and by state, not by individual students or by school or by sex. The objectives are listed in categories.

24. Percentage of achievement of an objective: the percentage of achievement is found by dividing the number of students who achieve the objective by the total number who were tested on the objective and multiplying by 100. The weighted percentages of achievement for the district and for the state were obtained in a similar fashion, but with the addition of appropriate weighting factors.

25. (25.1) Printed reports of the results are available.

(25.2) A preliminary statistical report is available for internal use. Individual reports are prepared for each school district. Annual public reporting of results is mandatory by state and by district.

(25.3) Objectives were broken down into classifications and subclassifications in order to present the objectives and results into meaningful groupings.

Definitions were provided for classifications and subclassifications as well as a sample item to indicate clearly how the skills were measured.

An interpretation of results for each subclassification is prepared by the state reading consultant. Tables present percentage of achievement on each objective for all districts and for the state, followed by a set of recommendations for improving second and fourth grade reading where interpretation found performance to be minimal or deficient. Eighty to one hundred per cent was generally considered satisfactory performance; 60-79% was generally minimal, and 0-50% generally was deficient. Minimal achievement denoted

areas in which certain weaknesses appear and developmental instruction should be given. Deficient areas needed corrective instruction utilizing varying instructional techniques and materials.

(25.4) We provide the media with reports and news releases—they often consider our data newsworthy enough to publish. Mass media cooperated in the release of information about the program prior to testing, such as adoption of goals, announcement of IOX contract for reading assessment materials, adoption of objectives for reading, and plans for the first February testing of 1971-72. This information was prepared by the Communications Media Service Center of the Department of Education. Several articles by department staff also appeared in Florida publications.

(25.5) No formal public meetings—testimony is given to legislative committees at their request.

26. Improvement in assessment materials and procedures to eliminate difficulties encountered in the 1971-72 and 1972-73 assessment.

Recommendations for improving second and fourth grade reading where needed both for the state and for each school district.

1972-73 data will be used to partially fulfill school accreditation requirements.

Provide technical assistance to districts in preparing local reports on local assessment results. (Related to section of law requiring local reports—a monograph is now in draft form.)

Program

27. September 1971: Items (first draft) delivered by item developer.

October 1971:	Internal item review.
November-December 1971:	External item review.
November-December 1971:	Printing of tests.
January 1972:	District coordinators received materials.
Late January 1972:	Schools received materials from district coordinators.
Early and mid-February 1972:	Assessment took place.
February 18, 1972:	Assessment materials returned to district coordinator.
February 25, 1972:	Assessment materials returned to state coordinator.
March 1, 1972:	Assessment materials received by Evaluation Section.
March 2-May 15, 1972:	Scoring (contracted out).
May 15-June 15, 1972	Final analysis.
June 15, 1972:	Preliminary report for internal use.
1972-73	
November 1972-January 1973:	A series of 3 public reports printed.

Basically the same schedule for 1972-73, but testing dates a little later.

28. $357,000 estimated SEA expenditures for operations; $95,000 estimated for objectives and exercises (this is a very rough guess); 4 people in other SEA units paid from other budgets—3 man-months each; 67 district coordinators paid from district funds—3 man-months each; 1,650 school coordinators and 300 to 400 school test administrators (many coordinators were also administrators) paid from district funds—1 man-week each. 50,000 students tested in 3 areas.

29. This was done by eliciting suggestions from school coordinators and test administrators, both by including a page for the purpose in the administration manual, and by holding a meeting for district coordinators to critique the program. The Department of Education also made staff visits to observe the testing procedure firsthand. The Evaluation Section examined its own activities for more efficient operation.

30. (30.1) Florida's Educational Accountability Act requires that a school district (each county) make a public report of assessment results by grade and subject area for each school in the district by 1973-74. This has caused considerable concern in the districts because of the sample size in the assessment program at the present time. Alternatives are presently being considered which will satisfy the requirements of the law.

(30.2) Accountability in Florida covers many areas in addition to statewide assessment. Areas other than assessment which contribute to Florida's accountability concept include school accreditation, cost analysis, instructional support systems, in-service teacher education, and comprehensive planning.

(30.3) The catalogues of objectives being developed by Florida's Research and Development activities are being made available to the districts, as they are completed, so that they may be integrated into their curriculum design and instructional programs.

Michigan Educational Assessment Program*

Purpose

1. The purposes of Michigan's Educational Assessment Program (MEAP) for 1973-74 are: (1) to determine whether fourth graders and seventh graders have mastered basic entry skills for those two grades, in reading and mathematics, and (2) to determine which school systems have the highest percentages of low achieving students, for purposes of awarding compensatory funds to those districts. State assessment is mandated by the legislature, and is part of a six-step accountability plan.

2. Goals of accountability and of assessment are published in a pamphlet titled *A Position Statement on Educational Accountability*, Michigan Department of Education, 1972.

3. 67 LEAs receive $200 per eligible pupil in compensatory funds, based on scoring low on state assessment tests. Schools must demonstrate student growth, in order to continue to receive the full $200. LEAs are expected to evaluate their own curricula and methodology in relation to assessment results, and to change their programs in an effort to improve student mastery of basic skills.

4. School administrators and teachers are the major audiences for assessment results. However, school district results are published for anyone who cares to use them. They have received considerable coverage in state newspapers.

Administrative Structure

5. (5.1) There is an Advisory Committee of eight persons, representing teachers, school counselors and psychologists, administrators, lay persons, curriculum specialists, researchers, and school public relations personnel. In addition, there is a larger Advisory Com-

*Prepared by Robert J. Huyser, Supervisor, Assessment Program, Department of Education.

mittee for state department research activities that also considers assessment problems. The Assessment Advisory Committee meets about once a month.

(5.2) There is a Director of State Assessment who reports to the Director of Research, Evaluation and Assessment Services. There are three other professionals on the assessment staff. There are four supporting staff.

(5.3) Ad-hoc advisers are used from time to time, for reviewing bids from potential contractors, and for advising on technical matters.

Instrumentation

6. The assessment program is primarily cognitive in nature. Assessment in the affective area has been in and out of the program.

7. For its first four years the tests used were of a normative type (similar to the *Sequential Tests of Educational Progress* of the Educational Testing Service). Beginning in the fall of 1973 the tests will be of a criterion-referenced type.

8. Reading and mathematics and language arts were assessed for the first four years. In 1973-74 it will be reading and mathematics. Future plans call for assessment in six other areas. They are health and physical skills, fine arts, foreign languages, science, social studies and music.

9. There are specific performance objectives in mathematics and reading and also in the other areas to be assessed in the future. They are (or will be) available from the Michigan Department of Education.

10. The objectives were developed by statewide committees of Michigan educators, under direction of state department personnel. Various professional organizations were involved (Michigan math teachers, music teachers, art teachers, etc.). These draft objectives were reviewed by broadly representative commissions organized in grade ranges.

11. Questions for the first three years were developed by the Educational Testing Service (ETS). In the fourth year, questions written by item development panels composed of Michigan subject matter specialists, teachers and others, were incorporated into the tests. Questions for the fifth year (1973-74) assessment were developed by staff members from four Michigan LEAs, under contract to the Department of Education. The LEAs were Pontiac, Waterford Township, Jackson, and Alma/St. Louis in consortium.

12. Questions are all multiple-choice.

13. Questions added in the fourth year (1972-73) were tried out in pre-tests given in the third year. The planning for and analysis of the results from the tryouts were handled by ETS. Questions for the 1973-74 assessment were tried out in the four LEAs where they were developed. The planning for and analysis of these tryouts were handled under a contract with CTB/McGraw-Hill.

14. Questions were selected for use in 1972-73 on the basis of judgmental reviews of the item development panels who had written the items. They made use of item analysis data from the prior year tryouts. Questions for use in 1973-74 were selected on the basis of judgmental reviews by department personnel and by CTB/McGraw Hill personnel and on the basis of the tryouts.

15. Questions are packaged into tests that are similar to conventional standardized tests. Separate answer sheets are used.

Data Collection

16. The total population is tested.

17. Michigan's assessment population is defined as all fourth and seventh graders enrolled in public schools in the state.

18. The only sampling done so far was for tryout of new items in the third year. Twenty-five pre-tests were each administered

to pupils in a different sample of districts, except Detroit. Schools were sampled in Detroit.

19. No sampling of items has been done except for pre-testing as above.

20. Data collection is by group administered, paper-and-pencil tests. Administrators are fourth grade teachers, each one testing his or her own class. At the seventh grade level, larger than class size groups are often tested by a counselor or principal.

Data Handling

21. Scoring is done by a contractor, who also prints the tests and manuals and distributes them.

22. Results are processed by the contractor who scores the tests.

23. Reporting categories are by grade with school and district summaries. Separate norms also are provided for district "type," e.g., metropolitan core city, cities, towns, urban fringe, and rural.

24. Reports have been made by raw scores, standard scores, percent below (pupil scores) percentile ranks (school and district means), and percentile bands. Types of scores to be used for 1973-74 will be pass/fail on each objective and percent of pupils who met (passed) an objective.

Reporting Results

25. (25.1) Printed reports of results for each school district are made available.

(25.2) They are sent to districts routinely. District results are made available some months later to anyone else who wants them, including the press.

(25.3) Interpretations are made by LEAs.

(25.4) The media is used to publicize results. News releases are prepared and a listing of district results is made available to the media.

(25.5) Public meetings are held in all parts of the state. The purposes are to acquaint school people and other with the purpose and methodology of assessment, to explain how to "read" the reports, and to answer questions about assessment.

26. There is no active utilization/application effort in relation to assessment results.

Program

27. In the first four years data collection took place in January, reporting of pupil results was in April and reporting of school and district results was in May. Explanatory publications accompanied the released results. In the 1973-74 program data collection will be October, pupil reporting will be in November and school and district reporting in December.

28. Total assessment costs for 1972-73, including assessment staff and contracting work, but not test development is estimated at $469,000. The per pupil cost (324,540 pupils) was $1.45.

29. There has been no systematic evaluation of the assessment program, but the change in direction for 1973-74 represents considerable thought of the assessment staff, other department personnel, the Superintendent of Public Instruction, and the State Board of Education. Both the Advisory Committee and other advisers have had considerable input into the department's continuing self-evaluation of the assessment program. Also, considerable input is received at public meetings.

30. Michigan is in a state of flux between its initial assessment program and the new, criterion-referenced thrust. The next major change in assessment will be the addition of testing

for school entry skills. Contracts for development and tryouts of materials will be let in 1973 in an effort to have them ready for assessment in the fall of 1974. No specific dates have been determined for the addition of other subject areas to assessment, or for an extension to high school.

GENERAL REFERENCES

Bard, Bernerd. "New York Accountability Plan Starts to Take Off." *Nations' Schools,* Vol. 90, September, 1972.

Berdie, Frances S. "What Test Questions Are Likely to Offend the General Public?" *Journal of Educational Measurement,* Vol. 8, No. 2, Summer, 1971.

Center for Statewide Educational Assessment. *CSEA Newsletter* (Princeton, New Jersey: Educational Testing Service).

Cooperative Accountability Project, *Characteristics of and Proposed Models for State Accountability Legislation* (Denver, Colorado: Colorado Department of Education, 1973).

Dyer, Henry S. and Solomon, Robert J. *Statewide Assessment: Its Future and Potential for Educational Reform.* Educational Testing Service, 1970. (Unpublished paper).

Educational Testing Service. *State Educational Assessment Programs* (Princeton, New Jersey: Educational Testing Service, 1971).

Finley, Carmen J. "National Assessment: Reports and Implications for School Districts." *The National Elementary Principal,* Vol. L, No. 3, January, 1971.

Finley, Carmen J. and Berdie, Frances S. *The National Assessment Approach to Exercise Development.* National Assessment Publications, July, 1970.

Grant, G. V. "Education's New Scoreboard." *Education Digest,* Vol. 38, January, 1973.

House, Ernest R. (ed.). *School Evaluation: The Politics and Process* (Berkeley, California: McCutchan Publishing Co., 1973).

Ludka, Arthur P. and Olson, Arthur R. "A Primer for Accountability in Education," *Issue II, Information Quarterly* (Denver, Colorado: Cooperative Accountability Project, March 1973).

Marland, Sidney, P. "Condition of Education in the Nation." *American Education,* Vol. 7, April, 1971.

Norris, Eleanor. "National Assessment: An Information Gathering and Information Dissemination Project." *Education,* April-May, 1971.

Norris, Eleanor. "What Are We Learning from the National Assessment?" *American Education,* Vol. 7, No. 6, July, 1971.

Olson, Arthur R. and Ludka, Arthur P. "Accountability: A Description as Viewed by the Cooperative Accountability Project," *Issue I, Information Quarterly* (Denver, Colorado: Cooperative Accountability Project, September 1972).

Olson, T. A. and Marvin L. "Evaluation: One State's Approach." *American Education,* Vol. 6, May, 1970.

Popham, James. *An Evaluation Guidebook.* Instructional Objectives Exchange, 1972.

Popham, James. "Procedural Guidelines: Developing IOX Objectives-Based Tests." *Technical Report No. 8,* Instructional Objectives Exchange, August, 1972.

Popham, James, *et al.* "State of the Art: Accountability in Education." *Journal of Educational Research,* Vol. 66, September, 1972.

Porter, John W. "Accountability Story in Michigan." *Phi Delta Kappan,* Vol. 54, October, 1972.

Rand Report, "How Effective Is Schooling: A Critical Review and Synthesis of Research Findings," 1972.

Stake, Robert E. "Language, Rationality, and Assessment." *Improving Educational Assessment and an Inventory of Measures of Affective Behavior* (Association for

Supervision and Curriculum Development, National Education Assessment, Washington, D.C., 1969).

Tyler, Ralph W. "National Assessment: A History and Sociology." *School and Society*, Vol. 98, December, 1970.

Tyler, Ralph W. "Testing for Accountability." *Education Digest*. Vol. 36, March, 1971.

Tyler, Ralph W. "What Is An Ideal Assessment Program?" Document prepared at meeting of special consultants on the State Testing Program, July, 1968 (Sacramento, California: California State Department of Education).

Wardrop, James L. "Towards a Broader Concept of Educational Assessment." Paper read at the Second Annual Mardi Gras Symposium on "Education: Instruction, Assessment, and Accountability," New Orleans, February 14, 1972 (Center for Instructional Research and Curriculum Evaluation, Univ. of Illinois, Urbana).

Womer, Frank B. and Mastie, Marjorie M. "How Will National Assessment Change American Education?" *Phi Delta Kappan*, Vol. 53, No. 2, October, 1971.

Womer, Frank B. "Research Toward National Assessment." Paper presented at the Western Regional Conference on Testing Problems, May, 1968.

Womer, Frank B. *What Is National Assessment?* National Assessment Publications, July 1, 1970.

Woodington, Donald D. "Accountability from the Viewpoint of a State Commissioner of Education." *Phi Delta Kappan*, Vol. 54, October, 1972.

Yee, Albert H. (ed.). *Perspectives on Management Systems Approaches in Education: A Symposium* (Englewood Cliffs, New Jersey: Educational Technology Publications, 1972).